Genetic Diversity, Population Structure, and Virulence Gene Polymorphisms in Nontypeable *Haemophilus influenzae*

by

Nathan C. LaCross

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Epidemiologic Science)
in The University of Michigan
2011

Doctoral Committee:

Professor Janet R. Gilsdorf, Co-Chair
Associate Professor Carl F. Marrs, Co-Chair
Professor Victor J. DiRita
Associate Professor Zhenhua Yang

ACKNOWLEDGMENTS

I would like to give my most sincere thanks and gratitude to my advisors and chairs of my committee, Drs. Gilsdorf and Marrs, for the incredible support and mentorship they've given over the years; the remainder of my committee, Drs. DiRita and Yang, for their valuable guidance; Drs. Rosenberg and Verdu for their exceedingly helpful advice on the analysis of population structure; the members of the Gilsdorf laboratory, particularly May Patel, for creating such a wonderful environment; my fellow Epidemiology students; and, of course, my family, friends, and loved ones, with the support of whom anything is possible.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ABBREVIATIONS

aa – amino acid

AIC - Akaike information criterion

bp – base pair

CI – confidence interval

DLV – double locus variant; used in eBURST

dNTP – deoxyribonucleotide

Fur – ferric uptake regulator

Hi – *Haemophilus influenzae*

Hib – type b *Haemophilus influenzae*

LOS - lipooligosaccharide

mAb – monoclonal antibody

MCMC – Markov Chain Monte Carlo

ME – middle ear

MLEE – multilocus enzyme electrophoresis

MLST – multilocus sequence typing

NAD – nicotinamide adenine dinucleotide

NTHi – nontypeable *Haemophilus influenzae*

OM – otitis media

OR – odds ratio

ORF – open reading frame

PCR – polymerase chain reaction

PFGE – pulsed field gel electrophoresis

PR – prevalence ratio

SSC – symmetric similarity coefficient

SLV - single locus variant; used in eBURST

SNP – single nucleotide polymorphism

ST – sequence type

Taq – *Thermus aquaticus*.

US – United States

GLOSSARY

Competence – ability of a cell to uptake exogenous DNA from its environment.

Crude – denotes values prior to adjustment for confounding variables.

Isolate – member of a bacterial species, not necessarily genetically distinct. Can (and often will) have the same genotype (i.e. be of the same strain) as other isolates. Frequently used interchangeably with 'strain' in the literature.

Strain – genetically distinct member of a bacterial species. Multiple isolates can be of the same strain. Frequently used interchangeably with 'isolate' in the literature.

V factor – nicotinamide adenine dinucleotide (NAD).

X factor – hemin.

ClonalFrame Parameters

δ - average tract length of a recombination event.

θ - mutational rate. Assumed to be constant on the branches of the topology.

ρ - recombination rate. Assumed to be constant of the branches of the topology.

ρ/θ - ratio of rates at which recombination and mutation occur, and therefore a measure of how often recombination events happen relative to mutations.

r/m - ratio of probabilities that a given site is altered through recombination versus mutation; a measure of how important the effect of recombination was relative to mutation.

structure Parameters

K - number of assumed populations.

Q - estimated individual membership proportion in each population *K.*

α - Dirichlet parameter for degree of admixture.

Chapter 1

Background and Public Health Significance

OTITIS MEDIA

Acute otitis media (OM) is an infection of the middle ear space, usually typified by fever, pain, and effusion. A predominantly childhood affliction, it is the most commonly diagnosed bacterial infection in children (6). By one year of age, 60% of children will have had at least one episode, and 17% of children will have had three or more.

The prognosis for the outcome of untreated otitis media is favorable, though the use of antibiotics in the management of OM is high. A Cochrane systematic review found that the symptoms of acute otitis media (though not necessarily the causative bacterial or viral infection) spontaneously resolved within two to seven days in 80% of children studied (42). Furthermore, there were no significant differences between antibiotic-treated children and untreated children in other clinical outcomes, such as tympanic membrane perforation or recurrence, and children treated with antibiotics were more likely to develop vomiting and diarrhea. Another study on antibiotic prescribing strategies for childhood otitis media came to similar conclusions, though they also found that immediate antibiotic treatment provided some symptomatic benefits after the first 24 hours (68). Indeed, in 2004 the American Academy of Pediatrics endorsed a

guideline that recommended initial observation rather than immediate antibiotic therapy for the management of acute otitis media in selected children (2).

Unfortunately, in rare cases untreated OM can lead to a number of serious complications, including perforation of the tympanic membrane, seizures due to prolonged high fever, mastoiditis, meningitis, brain abscesses, and death. Furthermore, many studies on the effectiveness of antimicrobial therapy in managing otitis media have substantial flaws, including imprecise diagnosing criteria, use of clinicians untrained in otoscopy, inadequate sample size, and ambiguous endpoints for cure and treatment failure (59, 142). Two recent studies sought to correct this problem, and performed randomized, blinded trials of the use of amoxicillin-clavulanate compared with placebo among children with a clear, certain diagnosis of OM in the age group at greatest risk (48, 130). As in prior studies, many children enrolled in the control groups improved without antibiotics, and more children in the treatment groups had associated adverse reactions. However, both research groups observed a significant, albeit modest, decrease in the duration of acute signs of illness among children who received the drug, indicating that antimicrobial therapy is likely beneficial for young children with accurately diagnosed acute otitis media.

Aside from the potential morbidity among children, OM has a considerable economic impact as well. A 1997 study found that the average total cost, including both direct (diagnosis and treatment) as well as indirect (missed wages, travel expenses, etc. for the parents) expenses, associated with a first OM infection was $107.81 ± 22.91. The direct and indirect costs associated with

recurrent OM infections were even higher, at $124.64 ± 40.24 per episode (56). Overall, the average annual total cost of otitis media infections has been estimated at approximately $3-5 billion in the United States (6, 56, 137). This figure is expected to rise as antibiotic resistance becomes more prevalent among bacteria that cause OM. Currently, OM infections account for 20% of all oral antibiotic prescriptions and greater than 50% of pediatric antibiotic prescriptions (6, 133, 137).

A variety of viral and bacterial pathogens can cause otitis media, including respiratory syncytial virus, rhinovirus, *Streptococcus pneumoniae*, *Moraxella catarrhalis*, and *H. influenzae*. Of the three major bacterial causative agents, *H. influenzae* (Hi) is responsible for 16-52% of infections, *S. pneumoniae* for 18-52% of infections, and *M. catarrhalis* for 11-23% of infections (7, 14, 29). The overwhelming majority of Hi strains that cause OM are nontypeable (NTHi), which differ from typeable strains in that they lack a polysaccharide capsule. Casey et al. found that between 1995 and 2003, there was a shift among the causative pathogens of OM in the US to a predominance of *Haemophilus influenzae* and a rise in the frequency of β-lactamase-producing *H. influenzae* isolated from OM patients (14). This may be due, in part, to the widespread use of the pneumococcal conjugate vaccine, which, while effective in lowering the incidence of otitis media caused by *S. pneumonia* serotypes included in the vaccine, does not protect against NTHi infection. Data have indicated that widespread use of the pneumococcal vaccine results in decreased pharyngeal colonization by *S. pneumoniae* and that NTHi may fill that niche with increased

colonization of the upper respiratory tract. (14, 30).

HAEMOPHILUS INFLUENZAE

Haemophilus influenzae are small, nonmotile, gram negative coccobacilli whose only natural hosts are humans (57). The species can be divided into two major groups differentiated by the presence or absence of a polysaccharide capsule. Six serotypes (a-f) have been identified among encapsulated strains (i.e. typeable strains, as the organisms agglutinate in the presence of type-specific antisera), each expressing structurally and antigenically distinct capsular polysaccharides (57). The most notorious Hi strains, expressing the type b capsule (Hib), typically cause invasive infections such as meningitis, bacteremia, and epiglottitis, and the capsule is an important pathogenic factor of the disease process (85). Infections with Hib have been nearly eliminated from developed countries following the introduction of an effective conjugate vaccine. However, the Hib vaccine provides no protection against infection with NTHi due to its lack of capsular antigen. Non-encapsulated (nontypeable, or NTHi) strains are much more frequent colonizers of the human pharynx, especially in children, and may cause a variety of respiratory infections, including otitis media, sinusitis, bronchitis, and pneumonia.

In addition to causing a range of respiratory (NTHi) and invasive (predominantly typeable strains, especially Hib) diseases, asymptomatic colonization of the nasopharynges is common, particularly in children. The carriage rate of NTHi among healthy children varies between 25-81%; this wide

distribution may be due to a number of factors, including proximity to other children (i.e. daycare centers), amount of antibiotic use, and exposure to secondhand smoke (8, 35, 128). Nasopharyngeal carriage of typeable Hi is considerably lower, from 3-7% (57).

The mechanisms by which some NTHi strains leave their commensal home in the nasopharynges to travel to the middle ear via the Eustachian tube are unknown. Once in the middle ear, however, these strains trigger an inflammatory response that results in acute otitis media. A number of host and environmental factors have been implicated in this process as well, including Eustachian tube dysfunction and obstruction, antecedent viral respiratory infection, allergies, exposure to cigarette smoke, and attending a daycare center (5, 6, 46). Furthermore, NTHi colonization is an active, dynamic process. A number of recent studies have shown that carriage is often marked by rapid turnover of strains as well as simultaneous colonization with multiple NTHI strains (22, 90, 134). Previous data gathered in our laboratory support these findings (35, 65, 128).

VIRULENCE FACTORS OF NONTYPEABLE *H. INFLUENZAE*

The role of the capsule as an essential virulence factor in Hib infections is well documented (85). Virulence in NTHi is less well understood, however, and no single gene or bacterial characteristic has been found to be universally associated with all disease strains. Moreover, whether NTHi strains isolated from diseased subjects (disease isolates) and those isolated from healthy

subjects (commensal isolates) differ in virulence potential remains unclear (28). This uncertainty is the focus of on-going research, and a number of adhesins (25), outer membrane proteins (46), and lipooligosaccharide (LOS) synthesis genes (103) have been shown to be more prevalent in disease isolates than in commensal isolates, suggesting a role for these loci in the disease process. Not all genes associated with virulence can be identified by prevalence studies, however. There are many genes that show no difference in prevalence between disease and commensal isolates (e.g. present in all or nearly all NTHi strains) that would thus yield a prevalence ratio close to one. These loci, such as *iga* and genes involved in the acquisition of iron, may still play a role in virulence.

Virtually all Hi strains secrete an endopeptidase, encoded by the *iga* gene, that specifically cleaves human IgA1 (57). Secretory IgA serves as an important mediator of immune protection at mucosal surfaces by binding and agglutinating pathogens, which results in steric hindrance of bacterial adhesin-epithelial cell binding (53). The IgA1 protease cleaves proline-serine or proline-threonine bonds in the hinge segment of the IgA1 α-heavy chain, separating the Fc and Fab fragments and ultimately resulting in hindrance of agglutination and bacterial clearance. While *iga* does not have a prominent role in this thesis, it does have an important function in our laboratory's working definition of what is 'true' NTHi amongst the overarching continuum of related strains and species. This function will be described in more detail in the following chapter.

Though the vast majority of NTHi express IgA1 protease, a study by Vitovski et al. found that strains isolated from patients with symptomatic

infections had significantly higher levels of IgA1 protease activity than strains isolated from asymptomatic carriers. Furthermore, they found differences in the size and sequence of the *iga* linker region, which connects the protease domain to the β-core autotranslocater (139). As the gene is present in both disease and commensal isolates, sequence polymorphisms may be responsible for the differences in protease activity. Further evidence of this relationship can be found in *Neisseria meningitidis*, where strains containing a mutant form of IgA1 protease lacking the usual consensus cleavage site in the linker region were recently described (140). Typically, self-cleavage at the consensus site is required for secretion of the mature extracellular form. The *N. meningitidis* mutants lacking the cleavage site were still able to secrete a functioning, mature protease, suggesting that the enzyme has the potential to cleave a wider range of proteins than previously suspected. Interestingly, Fernaays et al. recently identified a second IgA1 protease in NTHi, encoded by the *igaB* gene (38). The gene is transcribed, expressed, enzymatically active, and shows homology to the *iga* gene from *Neisseria* species.

Genes involved in the acquisition of iron and iron-containing molecules have also been shown to influence pathogenicity, despite many such loci exhibiting no differences in prevalence between disease and commensal isolates. The human host maintains concentrations of free iron considerably below the level needed for bacterial survival by sequestering it in hemoproteins such as transferrin, lactoferrin, haptoglobin, and hemoglobin (46, 144). Furthermore, as *H. influenzae* lack the enzymes necessary to synthesize the

heme precursor protoporphyrin IX, it has an absolute requirement for heme or heme derivatives for aerobic growth (145). This requirement is substantially reduced, though not eliminated, during anaerobic growth (57). Hi contains a number of systems to take up iron, usually in the form of transferrin, heme, hemoglobin, hemoglobin:haptoglobin complexes, and heme:hemopexin complexes.

The transferrin binding proteins, encoded by *tbp1* and *tbp2*, are outer membrane proteins responsible for Hi binding to transferrin and subsequent uptake of iron. Expression of both proteins is repressible by iron and under the control of the ferric uptake regulator (Fur). A study by Whitby et al. found that *tbp1* expression was over nine times greater during growth under iron/heme limiting conditions (144). Current understanding proposes that Tbp1 and Tbp2 act together to bind transferrin and remove iron, after which Tbp1 releases the iron into the periplasm by a TonB-dependent transport system (46).

Hi binding of heme and heme derivatives is accomplished by a number of independent pathways. The *hgp*'s are a group of four (*hgpA-D*) phase variable, CCAA nucleotide repeat-containing genes whose products are involved in the utilization of heme from both hemoglobin and hemoglobin:haptoglobin complexes. Different strains of NTHi contain between one and four *hgp* genes, and Morton et al. found that a mutant strain lacking all *hgp* genes had reduced virulence in a chinchilla model of otitis media as compared to wild type (79). Similarly, Seale and colleagues found that deleting all *hgp* genes from a type b strain reduced its (already limited) ability to initiate and sustain bacteremia in

weanling rats (121). Finally, a recent study from our laboratory found that *hgpB* was found in 100% of OM isolates versus 70% of commensal isolates, yielding a prevalence ratio of 1.44 and suggesting positive selection for *hgpB* containing organisms among OM-causing strains compared to commensal strains (146).

The *hxu* operon contains three genes, *hxuA-C,* that encode proteins involved in Hi binding to heme and heme:hemopexin. HxuA is secreted by Hi into the surrounding environment during growth and binds heme:hemopexin. HxuB is an outer membrane protein believed to facilitate the release of HxuA, and HxuC has similarity to TonB-dependent membrane proteins and could be involved with heme transport into the cell. Analysis of the operon suggests that *hxuA* and *hxuB* are transcribed as a unit, and both *hxuB* and *hxuC* have putative upstream Fur boxes, indicating they are also regulated by iron (46). Recently, Morton et al. constructed a mutant Hib strain lacking *hxuA-C* and found lower bacteremic titers and improved survival rates in five day old rats as compared with the wild type strain (82).

Less is known about two other iron acquisition genes, the hemin receptor *hemR* and the heme-binding lipoprotein *hbpA*. The *hemR* gene in *Yersinia enterocolitica* is capable of utilizing heme from a variety of heme-containing molecules (129), and, while similar studies have not yet been attempted in NTHi, our lab has found *hemR* to be 1.21 times more prevalent in OM strains than in commensal strains (present in 100% of OM strains and 83% of commensal strains) (146). Morton et al. recently demonstrated a role for *hbpA* in utilization of multiple heme sources, including heme, heme:hemopexin, hemoglobin, and

hemoglobin:haptoglobin (80). While further investigation is necessary for complete understanding of the relationship between iron acquisition and Hi virulence, the results described above suggest that a number of iron utilization genes may have a role. If certain alleles of these genes are better able to provide an NTHi strain with iron and heme, such a strain could have a selective advantage in normally privileged environments (i.e. the middle ear).

POPULATION STRUCTURE

Population structure, often used synonymously with the term population stratification in the literature, is the presence of systematic differences in allele frequencies between subpopulations within a population which are frequently due to differences in ancestry (1, 106). A frequent cause is nonrandom mating (or in bacteria, genetic exchange processes such as homologous recombination) between groups, followed by genetic drift of the allele frequencies in each group. This phenomenon tends to be very mutable over time, with new populations emerging and differences between other populations disappearing due to changes in gene flow. Genetic admixture can also result from such circumstances, in which individuals from different populations exchange genetic material, giving rise to populations and individuals with mixed ancestry.

Most methods for the characterization of population structure were developed for the study of human populations, and today generally make use of massive arrays of single nucleotide polymorphisms (SNPs) spread across the genome. Other markers frequently used include restriction fragment length

polymorphisms and microsatellites, which are repeating sequences of one to six base pairs of DNA. Many of the newest systems for genotyping humans, and thus providing information necessary for the discovery of population structure, combine upwards of two million of these markers on a single chip. While these systems are not typically available for the study of bacteria, the decreasing costs of obtaining large amounts of accurate DNA sequence data have enabled researchers to adapt methods used in human research to the study of population structure in microorganisms.

The characterization of population structure can be an end in and of itself, and provides useful information on both current and historical populations, as well as the forces that brought about the structure. For example, Kopelman et al. (60) recently confirmed that distant Jewish populations in different geographic areas largely share a common Middle Eastern ancestry that is still apparent despite varying degrees of admixture with more geographically adjacent non-Jewish populations, and Verdu et al. (138) found data in African Pygmy populations that suggest recent isolation, genetic drift, and heterogeneous admixture enabled rapid and substantial genetic differentiation from a unique ancestral population approximately 2,800 years ago. In bacteria, Falush et al. revealed that inferred ancestral populations of the human pathogen *Helicobacter pylori* can be mapped to historical human migrations (34). More recently, Sheppard et al. found evidence that two closely related zoonotic pathogenic species of *Campylobacter* are converging as a consequence of recent changes in gene flow (125).

Identifying population structure is also a vital component for most epidemiological association studies based on genetic data. These studies attempt to identify genetic markers that are statistically associated with an outcome of interest (often a disease). However, the presence of population structure in the data can cause substantial confounding in the form of spurious associations, which can result from both the marker and outcome frequencies varying across subpopulations (106, 107, 109). If the outcome of interest occurs at high frequency in a particular subpopulation, that group will be overrepresented among the cases and any marker that is at a higher frequency in that subpopulation than in the others will appear to be associated with the outcome. A number of methods have been developed to correct for population structure in association studies, including genomic control and structured association. In essence, genomic control methods use independent marker loci to adjust the distribution of standard test statistics (which are inflated due to the presence of population structure), while structured association methods infer the details of the population structure, assign individuals to clusters, and then stratify the test statistic by cluster. Ideally, both methods remove the confounding effect of population structure and permit the estimation of unbiased test statistics. These methods are well documented in the literature on association studies in humans (particularly the popular genome-wide association studies), but their adoption in the study of virulence- or disease-associated markers in bacteria is less thorough (32).

The research presented in this thesis examines the genetic diversity and

structure of nontypeable *Haemophilus influenzae*, as well as what may enable certain strains of the normally commensal bacteria to cause otitis media. Chapter 2 is an examination of the diversity present among, and relationships between, NTHi isolates collected from the throats of two healthy children attending a daycare center. Chapter 3 is an expansion of the methods presented in the prior chapter to a larger, randomly selected collection of both commensal and otitis media-associated isolates from three disparate geographic regions. Additionally, the phylogenetic relationships and population structure of this sample are explored. Finally, Chapter 4 utilizes this larger isolate collection to identify amino acid polymorphisms in the iron acquisition protein HemR that are associated with otitis media while accounting for the population structure identified in Chapter 3.

Chapter 2

Genetic Diversity of Nontypeable *Haemophilus influenzae*

INTRODUCTION

Microbial genomes exhibit an extremely wide range of diversity and plasticity. Some species, exemplified by *Mycobacterium tuberculosis*, are thought to be largely clonal with a fairly stable population structure (129). At the other extreme are species such as *H. pylori*, whose population structure is very non-clonal (70). Most bacterial species, including *H. influenzae*, fall somewhere in between and have some genomic regions that are relatively clonal and others that show evidence of a more diverse history.

A variety of different mechanisms play a role in the evolution of NTHi, including point mutation, deletions, hypermutability, recombination, and phase variation of protein expression by slipped strand mispairing of short tandem DNA repeats. Though all can act to alter the functional and antigenic nature of gene products, determining specific effects of recombination on bacterial virulence can be difficult. Horizontal transfer of DNA between strains and subsequent integration into the chromosome via homologous recombination can result in genomic 'mosaicism', or the mixing of genetic elements from multiple strains in a single chromosome. In most eukaryotic organisms, recombination makes a relatively small contribution to inherited polymorphisms, allowing patterns of

descent to be determined with reasonable clarity. With many species of bacteria, including Hi, recombination is a more frequent contributor to genetic change, and traditional phylogenetic methods can have difficulty reconciling the resulting patchwork of inherited genetic elements. Significant levels of recombination can obscure the phylogenetic signal (i.e. the tendency for related organisms to resemble one another) and cause phylogenetic trees constructed from different genes to disagree.

H. influenzae are naturally competent and possess dedicated machinery for the uptake and integration of exogenous DNA (112). A number of Hi genes show mosaicism, and thus evidence of recombination, in their sequences, including LOS synthesis genes, pilus genes, and the IgA1 protease (16, 17, 105). Furthermore, there is evidence of gene transfer across species lines, such as the probable transfer of the autotransporter *Iav* from *Haemophilus* to *Neisseria* and the tryptophanase gene cluster between *E. coli* and Hi (19, 72). Horizontal gene transfer, especially of genes potentially involved in virulence and interaction with the host, allows NTHi to rapidly adapt to its environment and dramatically increases the genetic diversity of the species.

Genetic diversity and population structure also differ between typeable and nontypeable isolates of *H. influenzae*. A number of studies have demonstrated that the population structures of *H. influenzae* types a-f are largely clonal. Musser et al. and Porras et al. analyzed many encapsulated isolates by multilocus enzyme electrophoresis (MLEE) and found evidence for a clonal pattern of descent (92-95, 104). In contrast, the population of NTHi is distinct

from that of typeable strains and appears to be both large and diverse (91, 104). One possible source for the greater genetic diversity seen among NTHi is an increased rate of recombination among those strains. A study by Pérez-Losada et al. found that nontypeable strains had a higher population recombination rate than typeable strains for four of the seven multilocus sequence typing (MLST) genes (102).

A number of recent studies have shown that NTHi colonization is an active, dynamic process. Dhooge et al. used arbitrarily primed PCR to type NTHi isolated from the nasopharynges of otitis-prone children and found that isolates collected more than four weeks apart from the same patient had different fingerprints, demonstrating rapid turnover of strains (22). A similar study by Samuelson et al. found that 58% of NTHi strains isolated from the nasopharynges of otitis-prone children were only found once, and 67% of strains had a minimum colonization period of two months or less (120). A number of other studies determined that children are often simultaneously colonized with multiple strains of NTHi. St. Sauver et al. and Farjo et al. found that a significant proportion of healthy children in daycare were colonized by multiple genetically distinct strains as determined by PFGE (35, 128); Murphy et al. found multiple NTHi strains in 26.3% of adults with chronic obstructive pulmonary disease (90); a study by Mukundan et al. revealed similar results by PFGE among four healthy adults (86).

To further explore the apparent diversity of NTHi that colonize the pharynges of healthy children, MLST was used to genotype 21 isolates of NTHi

from the throats of two healthy children attending the same daycare center. An additional 25 isolates from the same two children were found to be other species of *Haemophilus*. MLST, as opposed to more traditional molecular typing methods, such as pulsed field gel electrophoresis, provides unambiguous results that may be compared from laboratory to laboratory. The population recombination rate was estimated for each locus, and phylogenetic trees were constructed using the concatenated sequences of all seven loci.

MATERIALS AND METHODS

Bacterial Isolates

Nontypeable *Haemophilus influenzae* isolates from the throats of two healthy children (ID Nos. 22 and 26) attending the same daycare center and reported in a prior study (128) were used in the present study. Throat swabs had been collected once a week for three weeks, for a total of four swabs from each child. Samples from the two children were collected concurrently as described (128). In brief, samples were streaked on each of two chocolate agar plates supplemented with 300 µg bacitracin. Five colonies (selected for differing morphological features when possible) were selected from each plate, for a maximum of ten colonies for each child at each collection period. The isolates were frozen in sterile skim milk at -80˚ C.

Preparation of Genomic DNA

The stored isolates were grown overnight on chocolate agar plates (BD

Diagnostics, Sparks, MD) at 37°C with 5% CO_2 in a humid environment. Genomic DNA was isolated using the Wizard genomic DNA purification kit (Promega, Madison, WI.) according to the manufacturer's instructions, resuspended in distilled water, and stored at -20°C.

Species Determination

Requirement for X factor (heme) was determined by the porphyrin test developed by Lund and Blazevic (69), using the protocol of Farjo et al. (35). *H. influenzae* give a negative reaction while other *Haemophilus* species, including the common upper respiratory tract commensal *H. parainfluenzae*, have positive reactions. Horse blood agar plates (Remel, Lenexa, KS) were used with all isolates to test for hemolysis. Isolates were considered hemolytic if a distinct transparent zone was present in the agar surrounding the area of bacterial growth. The P6 outer membrane protein was characterized by immunoblot assay using the 7F3 monoclonal antibody (mAb) (kindly provided by Dr. Timothy Murphy) as described by Mukundan et al. (86). The 7F3 mAb binds to an epitope of P6 that is highly specific to *H. influenzae* (89).

The presence of *iga* was determined by PCR amplification of an 855 base pair (bp) conserved fragment (F: TGAATAACGAGGGGCAATATAAC; R: TCACCGCACTTAATCACTGAAT). PCR products were visualized on 1% agarose gels stained with ethidium bromide. The vast majority of *H. influenzae* strains contain *iga* that will amplify with these primers (or hybridize with a probe made from the PCR product), while the majority of *H. haemolyticus* strains do not

(55, 57, 86, 146).

Multilocus Sequence Typing

MLST was used to genotype the NTHi isolates as described by Meats et al. (77). PCR was used to amplify internal fragments of seven housekeeping genes, *adk, atpG, frdB, fucK, mdh, pgi,* and *recA*, in both the forward and reverse directions. The 25 µL reactions were carried out in either 96-well thin wall microtiter plates (MJ Research, Waltham, MA) or 0.2 mL Thermowell tubes (Corning Inc., Corning, NY) using a PTC-100 Thermal Cycler (MJ Research, Waltham, MA). The thermal cycler program followed the protocol of Meats et al. (77). The PCR products were purified using a QIAquick PCR Purification Kit (Qiagen, Valencia, CA), then submitted to the University of Michigan DNA Sequencing Core for DNA sequence analysis of both strands with the primers used for PCR amplification. The forward and reverse sequences were aligned, trimmed to the correct length, and the chromatograms checked for errors.

For each of the seven loci, the trimmed and edited sequences were compared to the database of *H. influenzae* sequences residing at the MLST website (http://haemophilus.mlst.net). Any sequences that did not match a previously established MLST allele were rechecked for accuracy, and then submitted to the curator of the MLST database for assignment of new allele numbers and entry into the database. Potentially new sequence types (STs) were also submitted to the MLST database.

Data Analysis

The population recombination rate ρ for each of the MLST genes was estimated following the methods of Pérez-Losada et al. (102) using a standard likelihood coalescent approach as implemented in the LDHat 2.1 program (76). LDHat also contains a likelihood permutation test (LPT) which was used to test the hypothesis of no recombination ($\rho = 0$).

The eBURST v3 program (http://eburst.mlst.net/) was used to examine the genetic relationships between NTHi isolates. eBURST uses MLST data to define a simple model of bacterial evolution in which an ancestral genotype, or founder, increases in frequency in the population, begins to diversify, and produces a cluster of closely-related genotypes known as a clonal complex (37). The most exclusive definition of a clonal complex was used, where STs are included in the group only if they share alleles at a minimum of six of the seven loci with at least one other ST in the group. Sequence types that differ at only one of the seven loci from another group member are known as single locus variants (SLVs), STs that differ at two of the seven loci from another group member are called double locus variants (DLVs), and so on. STs that differ at two or more alleles from every other ST in the dataset are known as singletons. The sequence type data of NTHi isolates from children 22 and 26, as well as all NTHi and type b *H. influenzae* isolates contained in the MLST database (accessed on November 15th, 2007), were analyzed as described by Feil et al. (37).

Further phylogenetic examination of the MLST loci were performed using the analyses of Erwin and colleagues (27). In a study of the genetic relatedness

of *Haemophilus influenzae* isolates using MLST, they constructed a maximum parsimony majority-rule consensus tree from all *H. influenzae* STs in the database (accessed September 18, 2006), including the nine STs identified in isolates from this study. The new technology algorithms implemented in the TNT version 1.1 program (43) were used to reconstruct 4,545 equally most parsimonious trees from the concatenated sequences of the seven MLST loci, using ST65 as the outgroup. These trees were used to generate the majority-rule consensus tree.

RESULTS

Phenotypic and Genetic Characterization

A total of 46 putative nontypeable *Haemophilus influenzae* isolates were included in this study: 18 from child 22 and 28 from child 26. During the initial isolation, X and V factor dependence was characterized by growth on media supplemented with disks impregnated with X and V factors (128). However, this method was subsequently found to have resulted in misclassification for a significant number of samples, as it is unable to distinguish *H. influenzae* from the highly related non-hemolytic, non-pathogenic *Haemophilus haemolyticus* species (88). Consequently, all isolates were characterized by porphyrin production, hemolysis on horse blood agar, presence of the *iga* gene by PCR, and reactivity to the 7F3 anti-P6 monoclonal antibody. For the purposes of this study, *H. influenzae* was defined as those isolates that possessed the *iga* gene, reacted to the 7F3 antibody, were non-hemolytic, and did not produce porphyrins

(i.e. required supplemental heme for growth). This is similar to the definition used by Murphy et al. to differentiate between *H. influenzae* and non-hemolytic variants of *H. haemolyticus* (88). The results of these tests are summarized in **Table 2.1**. Based on the above criteria, 21 of the 46 isolates were identified as *H. influenzae*, while the remaining 25 were determined to be other, related species within the genus.

MLST

Multilocus sequence typing was used to characterize the 21 *H. influenzae* isolates from children 22 and 26. A total of nine unique STs were identified from the *H. influenzae* isolated over the sampling period: three from child 22 and six from child 26. All three STs from child 22 and two STs from child 26 had been previously identified and entered into the MLST database. Seven of the nine STs differed from all other STs in this study by at least four of seven loci, and three STs differed by six of seven loci. This lack of shared alleles between isolates implies that the majority of STs collected are relatively unrelated. While STs 2 and 176 were present in samples collected sequentially on weeks one and two from child 22, the remaining STs were present during only a single sampling period. Furthermore, no ST was shared between the two children at any point, despite both attending the same daycare center.

Phylogenetic Analyses

Analysis of the MLST data using the eBURST v3 program was performed to

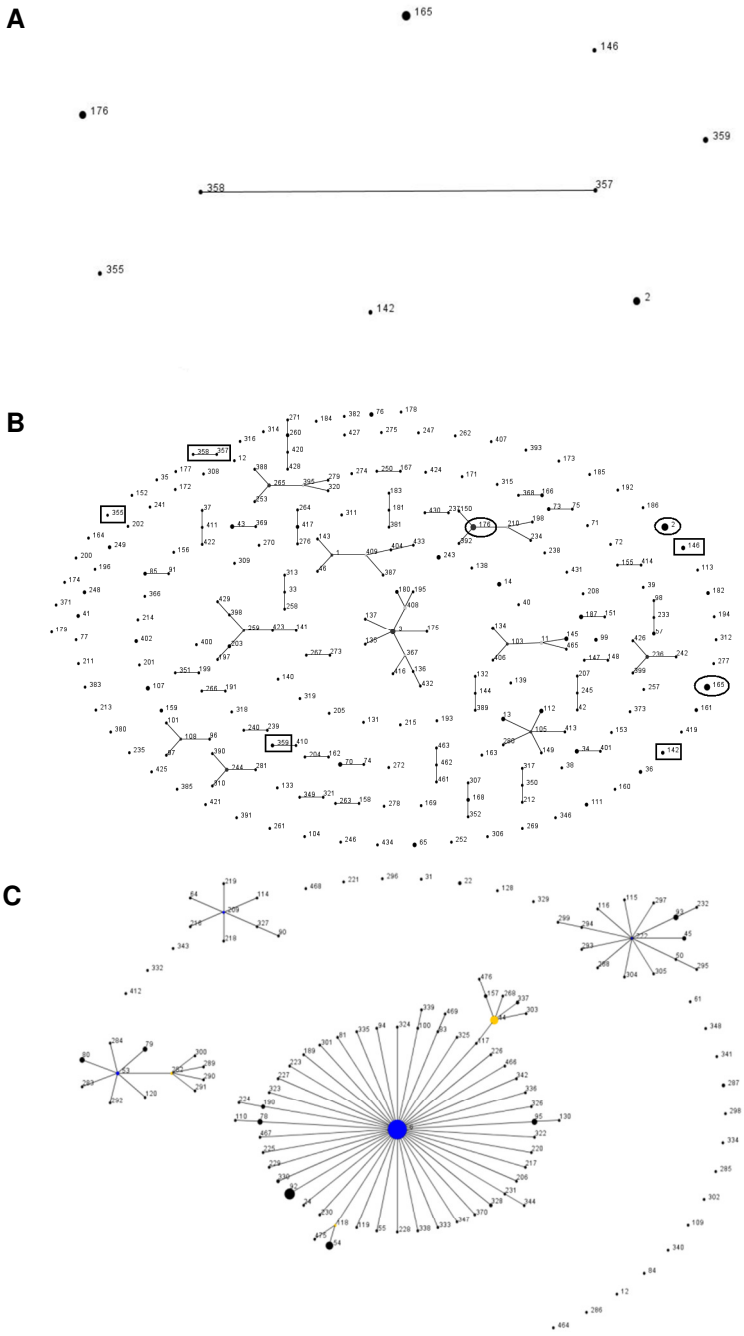
Table 2.1

Sample	Child 22						Child 26					
	Isolate	ST	*iga*	P6	Haemolysis	Porphyrin	Isolate	ST	*iga*	P6	Haemolysis	Porphyrin
1	22.1-21	176	+	+	-	-	26.1-21	359	+	+	-	-
	22.1-22	176	+	+	-	-	26.1-22	359	+	+	-	-
	22.1-23	176	+	+	-	-	26.1-23	355	+	+	-	-
	22.1-24	2	+	+	-	-						
	22.1-25	NA	-	-	+	+						
2	22.2-21	176	+	+	-	-	26.2-25	142	+	+	-	-
	22.2-22	2	+	+	-	-	26.2-21	NA	-	+	-	-
	22.2-23	2	+	+	-	-	26.2-22	NA	-	+	-	-
	22.2-25	2	+	+	-	-	26.2-23	NA	-	+	-	-
	22.2-24	NA	-	-	+	-	26.2-24	NA	-	+	-	-
							26.2-26	NA	-	+	-	-
							26.2-27	NA	-	+	-	-
							26.2-28	NA	-	+	-	-
3	22.3-21	NA	-	-	+	+	26.3-23	357	+	+	-	-
	22.3-22	NA	-	-	-	+	26.3-27	358	+	+	-	-
							26.3-22	NA	+	-	+	-
							26.3-24	NA	-	+	-	-
							26.3-28	NA	-	+	-	-
							26.3-25	NA	-	-	-	-
							26.3-26	NA	-	-	-	-
4	22.4-21	165	+	+	-	-	26.4-24	146	+	+	-	-
	22.4-22	165	+	+	-	-	26.4-21	NA	-	+	-	-
	22.4-23	165	+	+	-	-	26.4-22	NA	-	+	-	-
	22.4-24	165	+	+	-	-	26.4-29	NA	-	+	-	-
	22.4-25	165	+	+	-	-	26.4-210	NA	-	-	+	-
	22.4-26	165	+	+	-	-	26.4-23	NA	-	-	-	-
							26.4-26	NA	-	-	-	-
							26.4-25	NA	-	-	-	-
							26.4-27	NA	-	-	-	-
							26.4-28	NA	-	-	-	-

Characteristics of *Haemophilus* isolates. Sequence types, as determined by MLST, were assigned to those isolates with characteristics indicative of *H. influenzae*, as follows: *iga*, presence (+) of the *iga* gene as determined by PCR; P6, reactivity (+) to the *H. influenzae*-specific 7F3 mAb epitope; Hemolysis, hemolytic activity (-) during growth on horse blood agar plates; Porphyrin, presence (-) of fluorescence under UV light after incubation with δ-aminolevulinic acid. NA, not applicable.

assess the relative relatedness among NTHi isolates isolated from the two children. Sequence types 357 and 358 from child 26 are SLVs and form a group, while the remaining STs are singletons (**Figure 2.1a**). This pattern mirrors that seen in an eBURST analysis of all NTHi isolates residing in the MLST database, which shows seven groups of five or more STs and a large number of STs that are at least double locus variants of all other STs (i.e. they differ by two or more alleles from every other ST in the dataset) (**Figure 2.1b**). The pattern shown by typeable STs, however, is considerably different. eBURST analysis of all type b isolates in the MLST database reveals that one large clonal complex predominates, surrounded by three smaller complexes and a number of singletons (**Figure 2.1c**). Like the NTHi STs residing in the MLST database, the type b STs represent a very wide range of isolation dates and geographic locations. Overall, this suggests that in contrast to type b strains, NTHi are considerably more diverse and show a less clonal pattern of descent, as has been suggested in other studies using different analytic techniques (94, 119). Erwin et al. (27) constructed a maximum parsimony majority-rule consensus tree from 4,545 equally most parsimonious trees using the concatenated MLST sequences from all *H. influenzae* sequence types in the MLST database, including the nine STs found in the two children in this study. These nine NTHi STs are scattered in the tree, implying that they are more related to other STs within the MLST database than to each other (**Figure 2.2**).

Figure 2.1



eBURST analysis of selected *H. influenzae* isolates. STs differing from another ST by only one locus are connected by lines and form a group. The size of the circles is proportional to the abundance of the corresponding STs in the data set, and the relative placement of unconnected STs is random. Blue circles represent clonal complex founders, while yellow circles are subgroup founders. **A**. The nine STs from children 22 and 26. **B**. All NTHi STs in the MLST database (accessed 11-15-2007. STs from child 22 are circled, while STs from child 26 are in squares). **C**. All type b STs in the MLST database (accessed 11-15-2007).

Recombination

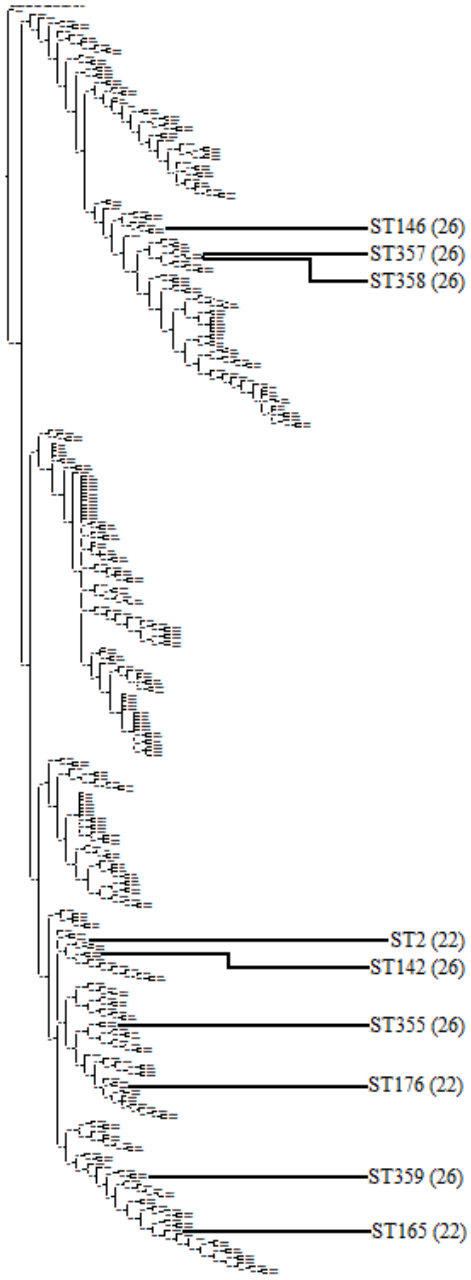
The population recombination rate, ρ, of each of the seven MLST loci from the NTHi isolates from the two children was determined using the standard likelihood coalescent approach implemented by the LDHat 2.1 program (76). Under this model, ρ can be expressed as $8N_ect$ for haploid organisms, where N_e is the effective population size, c is the per base rate of initiation of gene conversion, and t is the average gene conversion tract length, all of which were estimated from the data. Estimates of ρ varied considerably across the different loci, ranging from a low of zero for the *fucK* and *recA* genes to a high of 94 for the *mdh* gene (**Table 2.2**). Significant evidence of recombination was observed at the α = 0.05 level for two of the seven loci (*frdB* and *pgi*) using the likelihood permutation test contained within LDHat. Two additional loci, *adk* and *mdh*, have p-values of 0.106 and 0.096 respectively, showing a strong trend toward significance, especially considering the small number of alleles analyzed. These data are similar to the values found by Pérez-Losada et al. when analyzing all *H. influenzae* isolates in the MLST database (as of January 2004), as well as data

Table 2.2

Gene	Recombination Rate (ρ)[a]	p-value[b]
adk	6	0.106
atpG	9	0.516
frdB	14	0.000
fucK	0	0.896
mdh	94	0.096
pgi	59	0.000
recA	0	0.000

Population recombination rate for each MLST locus.
[a] population recombination rate estimated by the LDhat software.
[b] p-values estimated using the likelihood permutation test in the LDhat software.

Figure 2.2



Maximum parsimony majority-rule consensus tree constructed by Erwin et al. (27) from 4,545 equally most parsimonious trees, using all *H. influenzae* STs in the MLST database. The nine STs identified in this study are marked, and the number in parentheses indicates the child from which the STs were collected.

subsets consisting of strains used by Meats et al. (77) in their development of the *H. influenzae* MLST scheme (102). Even with the small number of sequences analyzed here, it is clear that the rate of recombination is not uniform across the NTHi genome.

DISCUSSION

Recognition of the wide range of diversity and plasticity exhibited by microbial genomes represents a major paradigm shift in microbiology. Definition of inter- and intra-species genetic diversity among bacteria has been greatly facilitated by recent whole bacterial genome sequencing projects. In addition to strain Rd KW20 (a type d isolate that lost capsule expression) (39), nine *H. influenzae* genomes are listed as having been fully sequenced, with a further 21 in the sequencing or assembly stages on the NCBI genome project web page for Hi. At

least three of these strains are commensal isolates gathered from healthy individuals. Studies by Harrison et al., Munson et al., and Shen et al. compared some of these genome sequences to explore *H. influenzae* diversity and to identify genetic regions specific to NTHi (47, 87, 123).

The advent of nucleotide sequence based genotyping methods has also aided the study of bacterial diversity. While MLST samples only a minute fraction of the genome (0.17% in strain Rd KW20) and provides no information regarding differential gene content or arrangement, it is easier and less resource intensive than full genome sequencing and provides useful information beyond that available from many prior techniques like pulsed field electrophoresis (PFGE), such as analyses of sequence divergence, recombination, phylogenetics, and population structure (36). Previous studies have demonstrated the utility of MLST in assessing diversity and relationships between isolates of *H. influenzae*. Erwin et al. used MLST to type a number of commensal, otitis media-associated, and invasive NTHi isolates and, using a UPGMA dendrogram, demonstrated that eight of eleven biotype V isolates formed a more distinct cluster than other isolates, and that isolates containing *hmw*-related sequence were separated from those lacking *hmw* (26). In a comparison of MLST to a typing strategy based on 16s rRNA sequence, Sacchi et al. found that both methods corroborate the clonal nature of typeable *H. influenzae* and the lack of clonality of NTHi (119).

This study used MLST to further explore the genetic diversity of NTHi that colonize the pharynges of two healthy children. The diversity seen by MLST corroborates that seen by pulsed field gel analysis of the same isolates (128),

which assesses relationships based on the presence and genomic positions of the DNA sequences digested by specific restriction enzymes. This corroboration strengthens and broadens the degree of the NTHi diversity seen within two individuals.

Despite prior identification as NTHi by standard laboratory techniques, 13 of the 46 total isolates from the two children were found to be non-*influenzae* species of *Haemophilus*. These isolates were hemolytic, did not react with antibody directed against the *H. influenzae* specific epitope on P6, and/or produced porphyrins, all of which are characteristics not associated with *H. influenzae*. A further 12 isolates reacted with the *H. influenzae* P6 7F3 mAb but were negative for the *H. influenzae iga* gene by PCR, the presence of which has been considered a defining feature of *H. influenzae* (57). Recent work has found that 20-40% of apparent *H. influenzae* isolates isolated from healthy, colonized individuals are in fact non-hemolytic variants of the closely related, but non-pathogenic, species *H. haemolyticus*, and that standard laboratory techniques are inadequate to distinguish these from *H. influenzae* (88). It is likely that at least some of the 25 non-*H. influenzae* isolates identified in this study are non-hemolytic *H. haemolyticus*.

The 12 *iga* negative, 7F3 antibody reactive isolates present an interesting situation, as they react with the ostensibly *H. influenzae* specific mAb but cluster with *H. haemolyticus* and other non-*H. influenzae* species in maximum parsimony cladograms constructed from the concatenated sequences of a subset of the MLST genes that are present in each species (*adk, pgi,* and *recA)*

(data not shown). These isolates may represent strains that are intermediate between *H. influenzae* and related species, which may not be surprising given the relatively high levels of recombination observed here and in other studies (17, 102). This phenomenon has been previously described among *Neisseria* species, where Hanage et al. found that some strains of *N. mucosa*, *N. sicca* and *N. subflava* could alternatively be sister to *N. meningitidis* or *N. lactamica* depending on where the tree is rooted, while other strains of the same species were basal to both *N. meningitidis* and *N. lactamica* (45). Similarly, Mukundan et al. identified a commensal *Haemophilus* strain that did not fit a strict definition of either *H. influenzae* or *H. haemolyticus* (86). These data emphasize the idea that conventional species definitions for bacteria are in some sense arbitrary constructs that do not fully take into account the continuum of characteristics found among members of closely related species.

Three sequence types were found among the 14 NTHi isolates from child 22, while six STs, including four previously undescribed STs, were identified from the seven NTHi isolates from child 26. One of the STs from child 22 and all six of the STs from child 26 were found only during a single sample period. Furthermore, no ST was shared between the two children at any point in the study despite attending the same daycare center, though other children from the center did exhibit strain sharing (128). Seven of the nine STs differ from all other STs in this study by at least four of the seven loci, and four STs differ by six loci. eBURST analysis revealed no major clonal complexes, consistent with analysis performed on all NTHi STs in the MLST database. The maximum

parsimony majority-rule consensus tree constructed by Erwin and colleagues (27) showed that the nine STs are scattered in the tree and do not form any major clades. These data suggest the presence of a large pool of unique NTHi strains within the children sampled and inadequacy of the bacterial isolation technique to detect all NTHi STs present with each child during a single sampling procedure. Alternatively, this larger pool of unique NTHi may exist within the community and the lack of persistence and relatedness of the collected isolates reflects the rapid turnover of strains within each child. A third, though relatively unlikely, option is that the high diversity observed in this study is the result of extremely rapid evolution. The true scenario may be a combination of these possibilities, in which individuals quickly gain and lose rapidly evolving NTHi strains, and standard collection techniques are incapable of providing an accurate survey of the genetic differences present among NTHi residing within the nasopharynges.

Chapter 3

Population Structure of Nontypeable *Haemophilus influenzae*

INTRODUCTION

Population structure, the presence of systematic differences in ancestry between subgroups of a population, has been a focus of research in eukaryotes, particularly humans, for decades, and is increasingly being surveyed in bacterial species. Most early studies of bacterial population genetics, diversity, and structure used the then standard eukaryotic technique of MLEE. Selander and colleagues were some of the early pioneers in applying this method to a number of bacterial species, including *Escherichia coli*, *H. influenzae*, *Legionella pneumophila*, and species within the genus *Bordetella* (122). Many of the species studied exhibited what were thought to be surprising levels of clonality for recombining organisms (1). Later publications, particularly by Smith et al. (126, 127), clarified and expanded on these ideas, remarking that in some species, instead of true clonality, it may be that recombination is so frequent that apparently clonal groups might exist only temporarily and would dissipate once sufficient recombination occurred. In other cases, such as in *N. meningitidis*, rapid epidemic spread of particular genotypes may result in apparent clonality, particularly when most isolates under study are isolated from cases of disease. Apparent clonal population structure can also arise when the sample analyzed

consists of a mixture of different populations, and recombination is frequent within each population but rare between them, as in *Rhizobium meliloti* (127).

The introduction of MLST in 1998 rectified some of the disadvantages of MLEE (71). Unlike MLEE, which defines electrophoretic types based on the mobility of metabolic enzymes when run on a gel, MLST data are based on DNA sequence and thus is readily portable and comparable between laboratories. Use of MLST typing has allowed the collection and dissemination of immense amounts of genetic data on dozens of bacterial species, all stored in publicly available databases. Equally important is that the sequence data generated by MLST genotyping is far more useful for investigation of population genetics, diversity, and structure, and allows the adoption of sequence based methods originally designed for other organisms.

The mechanisms underlying the high genetic diversity observed among NTHi in the prior chapter influence population structure of the organism. While the population structure of *H. influenzae* has been investigated in previous studies, most of the literature dates back to the MLEE era and principally used typeable strains. In 1985, Musser and colleagues characterized 177 type b *H. influenzae* isolates recovered from children with invasive disease by MLEE and determined that the sample exhibited significant clonality and major differences in the genetic structure of populations from the United States (US) and the Netherlands (92). Similarly, a study of over 2,200 typeable isolates from 30 countries was characterized by MLEE and again found that the population structure overall was clonal with strong patterns of geographic variation and a

limited number of evolutionary lineages that largely corresponded to serotype (93, 94). For example, Hib isolates of electrophoretic type 100 comprised 4.5% of their Canadian sample, but this genotype was not found among isolates from the US. However, like many studies of clinically significant microorganisms, the vast majority of isolates were collected from cases of disease; of the 2,209 isolates in the study, nearly 90% were serotype b, and less than 5% of these were obtained from healthy carriers.

Interestingly, the population structure of *H. influenzae* appears to differ between typeable and non-typeable strains. In a 1985 study of 242 Hi disease isolates (65 nontypeable and 177 type b), Musser et al. found that each NTHi isolate was of a unique electrophoretic type and that none of these were shared with a Hib isolate, indicating that the population of NTHi is extremely heterogeneous and distinct from that of typeable strains (91). Porras et al. used MLEE to characterize 135 Hi isolates from Sweden and the US, of which 81 were nontypeable. Seventy electrophoretic types were identified among the 81 NTHi isolates, and no electrophoretic type was shared between the geographic regions (104). While five electrophoretic types were found in both nontypeable and type b isolates, their version of MLEE assayed only six enzymes compared to the 15 used by Musser et al. (91), drastically reducing the discriminatory power of their method. More recently, Erwin et al. examined the population structure of all 656 Hi isolates in the MLST database as of 2006, including 322 NTHi isolates, based on a maximum-parsimony analysis (27). However, as the authors noted, there may have been sampling bias, as nearly 90% of the NTHi in the MLST database

were isolated from patients with symptomatic infections. Furthermore, NTHi submitted to the MLST database come from many different researchers and do not conform to any specific sampling scheme. Nevertheless, they were able to identify well defined phylogenetic groups of NTHi that differed in genetic content, though there was little apparent clustering by geography or clinical site of isolation.

In this study, the genetic diversity, phylogenetic relationships, and population structure of nontypeable *H. influenzae* was explored. A diverse collection of 170 commensal and otitis media-associated isolates from three disparate geographic regions were genotyped by MLST. Interestingly, 15 isolates (14 commensal and one disease-associated) were found to have a deletion of *fucK*, one of the MLST loci. Genetic diversity and phylogenetic relationships between the isolates were assessed using eBURST and the ClonalFrame program, while population structure was characterized using *structure*.

MATERIALS AND METHODS

Bacterial Isolates

An initial set of 204 putative NTHi isolates was selected from existing collections representing three distinct geographic regions (Finland, Israel, and the US). Within each geographic region, half of the isolates had been collected from the middle ears of children with acute otitis media (hereafter designated ‘OM isolates’) and the remaining half had been collected from the throats or

nasopharynges of healthy children (commensal isolates). This yielded six subgroups of isolates, the majority of which have been previously described in the literature: Finland OM (58); Finland commensal (136); Israel OM (67); Israel commensal (44); US OM (61), as well as unpublished isolates from Dr. Stan Block and Dr. Alejandro Hoberman; and US commensal (35, 128). Sixty total isolates from Finland and Israel and 84 total isolates from the US were randomly selected from within each subgroup for inclusion in the study. Only a single isolate was selected from each child, and all isolates were collected within an eight year period (1994 – 2002) from children under seven years of age (**Table 3.1**). The isolates were frozen in sterile skim milk at -80˚ C for storage.

Preparation of Genomic DNA

The stored isolates were grown overnight on chocolate agar plates (BD Diagnostics, Sparks, MD) at 37℃ with 5% CO_2 in a humid environment. Genomic DNA was isolated using the Wizard genomic DNA purification kit (Promega, Madison, WI.) according to the manufacturer's instructions and resuspended in 1x Tris-EDTA buffer (10 mM Tris-HCL and 1mM EDTA at pH 8). The majority was kept at -20℃ for storage, with a small aliquot stored at 4℃ for use.

Isolate Exclusion Criteria

All selected isolates were tested to ensure their identity as NTHi by the methods described in the subsequent two sections. Isolates were excluded from

the final dataset if they met any of the following criteria: missing from the isolate collections or unable to grow under standard conditions, presence of the capsule locus genes, identification as non-Hi, or persistent and unresolvable contamination (e.g. superimposed peaks in the sequence chromatograms not resolvable by multiple re-isolations of genomic DNA from single colonies). A total of 34 isolates met the exclusionary criteria, leaving 170 NTHi isolates in the final dataset.

MLST

MLST was used to genotype the NTHi isolates following the protocol of Meats et al. (77) and as described in the previous chapter, with the exception that a 170-9701EDU MyCycler thermal cycler (Bio-Rad, Hercules, CA) was used.

In most cases, strains closely related to, but not part of, Hi are negative for the MLST locus *fucK*, and this difference has been exploited to distinguish between the two groups (97, 99). However, a recent paper by Ridderberg et al. has described at least one apparent *H. influenzae* strain in which the entire six gene fucose operon (of which *fucK* is a member) is missing by PCR analysis, indicating that this absence is not a reliable indicator of species identity (113). In the present study, the PCR techniques described by Norskov-Lauritsen were used to determine the presence or absence of the fucose operon in every isolate in which reliable amplification of *fucK* could not be achieved. Briefly, PCR was conducted with primers flanking the fucose operon using LongAmp *Taq* DNA polymerase (NEB, Ipswitch, MA), and the products were visualized by agarose

gel electrophoresis stained with ethidium bromide. Isolates containing the operon yielded amplicons of approximately 10 kb, while isolates lacking the operon had amplicons of approximately 2 kb.

Determination of Capsule Typeability

All selected isolates underwent testing to ensure their identity as true nontypeable *H. influenzae*. The highly conserved *bexA* and *bexB* genes, which are required by typeable strains for the transport of capsule components across the outer membrane, were assayed by PCR following the protocol of Davis et al. (18). The advantage of this method over traditional slide agglutination techniques using type-specific antisera or methods detecting *bexA* alone is that *bexB* PCR will detect rare strains that are *bexA* negative but *bexB* positive, which renders them phenotypically nontypeable but genetically far closer to typeable strains. A positive result for either gene indicates the presence of the *cap* locus and thus that the isolate is at least genetically typeable; such isolates were excluded from the final dataset. The capsule type was identified by PCR of the capsule type specific regions of the *cap* locus in isolates that were *bexA* and/or *bexB* positive using the method of Falla et al. (31).

eBURST

The eBURST v3 program was used to analyze the MLST data from the final dataset of NTHi isolates as described in the previous chapter. A paper by Turner et al. described a method of assessing the reliability of the clonal complex

assignments inferred by eBURST by calculating the proportion of STs in the largest clonal complex identified by the program (135). If the largest clonal complex contains greater than 25% of the STs in the sample, they judged that the performance of the program is likely to be suboptimal. Reliability was assessed by this method using all NTHi STs in the MLST database (accessed March 30th, 2011).

Phylogenetic Analysis

As demonstrated in the prior chapter, the traditional methods of distinguishing between species within the genus *Haemophilus* are inadequate when faced with non-hemolytic variants of *H. haemolyticus*, as the traditional differentiating factor between *H. influenzae* and *H. haemolyticus* is hemolytic activity. McCrea and colleagues determined that a phylogeny constructed from multiple loci accurately distinguished known *H. influenzae* strains from known *H. haemolyticus* strains in a panel of 197 isolates (75). A similar method was employed here using ClonalFrame 1.1 (24). This program is based on a neutral coalescent model of genetic diversification that estimates the clonal relationships between members of a bacterial population while accounting for the way recombination occurs in such populations, thus allowing more accurate phylogenetic inferences. ClonalFrame is well suited to the analysis of MLST data, and each locus is assumed to be independent and distant from the previous one so that a given recombination event only affects a single locus (23). The program also infers when and where recombination events took place in the

evolutionary history of the sample and estimates population-wide evolutionary parameters, such as the mutation and recombination rates.

Sequence data from six of the seven MLST loci (excluding *fucK,* as most non-Hi and some true Hi are missing the operon) from a dataset of 181 of the original 204 selected isolates was analyzed with ClonalFrame. Excluded were five missing isolates, eight typeable isolates, four persistently contaminated isolates, and six (five US commensal, one US OM) isolates confirmed as non-Hi (75). In addition, three non-Hi strains from the *Pasteurellaceae* family were included: *H. haemolyticus* strain HK386 (98), *H. parainfluenzae* strain T3T1 (GenBank ID FQ312002.1), and *Pasteurella multocida* strain Pm70 (74). Two independent ClonalFrame runs of 200,000 iterations each were performed, using default values for all options. The first 100,000 iterations were considered the burnin period and were discarded, and the remaining iterations were sampled every 100 generations to produce 1,000 topologies in the posterior sample. Convergence of the Markov Chain Monte Carlo (MCMC) was assessed by the Gelman-Rubin test (40) as implemented by ClonalFrame. A Gelman-Rubin statistic above 1.2 indicates poor convergence (23); the statistics for all parameters were below this value when convergence was compared between the two runs. An unrooted majority consensus tree was constructed from the posterior sample using SplitsTree 4.11.3 (49). Isolates that clustered with *H. haemolyticus, H. parainfluenzae,* or *P. multocida* were considered to be non-NTHi and excluded from further analysis.

To assess the phylogenetic relationships between isolates deemed to be

true NTHi, the final dataset of 170 isolates was input into ClonalFrame. All seven MLST loci were used, with *fucK* negative isolates treated as having a gap at that locus. The methods used were identical to those described above, with the exception that 400,000 total iterations (200,000 burnin iterations and 200,000 sampling iterations) were performed in each of two independent runs, yielding 2,000 topologies in the posterior sample.

Population Structure

The *structure* 2.3.3 program was used to identify and assess population structure in this collection of NTHi using MLST data (108). *Structure* is a model-based clustering method that has been extensively used to infer population structure in humans (52, 60, 138, 143), animals (114, 116), plants (13, 50), and bacteria (21, 34, 124, 125). It assumes a model in which there are *K* populations (where *K* may be unknown), each of which is characterized by a set of allele frequencies at each locus. Individuals in the sample are probabilistically assigned to populations in the no-admixture model, or jointly to two or more populations if their genotypes indicate admixture under models that allow such phenomena. A key assumption is that loci are at Hardy-Weinberg equilibrium and linkage equilibrium within populations, and individuals are assigned to populations to achieve this. *Structure* also assumes that loci are independent within populations, which is likely to be violated when using sequence data as in this study. However, if the sequence data is from multiple independent regions (e.g. MLST data), *structure* may still perform well as long as there is enough

independence across regions that linkage disequilibrium *within* regions does not dominate the data. The main cost of dependence within regions is that *structure* will underestimate the uncertainty in the assignment of particular individuals (110). Another potential model violation is including multiple family members, which in some instances can lead to overestimation of *K*. However, this tends to have little effect on the assignment of individuals to populations for a given value of *K* (110).

In this study, *structure* was run on two datasets: the first contained all isolates, while the second consisted of only one example of each unique ST found in the sample. Analysis of population structure was performed on the two datasets to assess the impact of including multiple isolates of the same genotype. As some isolates were missing the *fucK* locus, only the remaining six MLST loci were used in both datasets. Twenty replicate runs of 100,000 burnin iterations and 100,000 sampling iterations were performed for each value of *K*; all were based on the admixture model with correlated allele frequencies and independent values of the Dirichlet parameter α for each assumed population *K*. Convergence was assessed by visually examining the parameter traces. The number of assumed populations was increased until adding a population became uninformative (i.e. the individual membership proportions (Q) for the added population were very low and/or few individuals had a large portion of their ancestry from that population). The *Greedy* algorithm implemented in CLUMPP 1.1.2 (51) was used to identify potential distinct modes among the 20 replicate runs for each *K* value. To be within the same mode, two replicate runs at a given

K must have had a symmetric similarity coefficient (SSC) ≥ 0.9. The estimated individual membership proportions were then averaged among all runs within the same mode for a given value of *K*. Plots of *structure* results were produced using *distruct* 1.1 (115). This method of analysis is similar to those described in earlier studies (60, 138, 143).

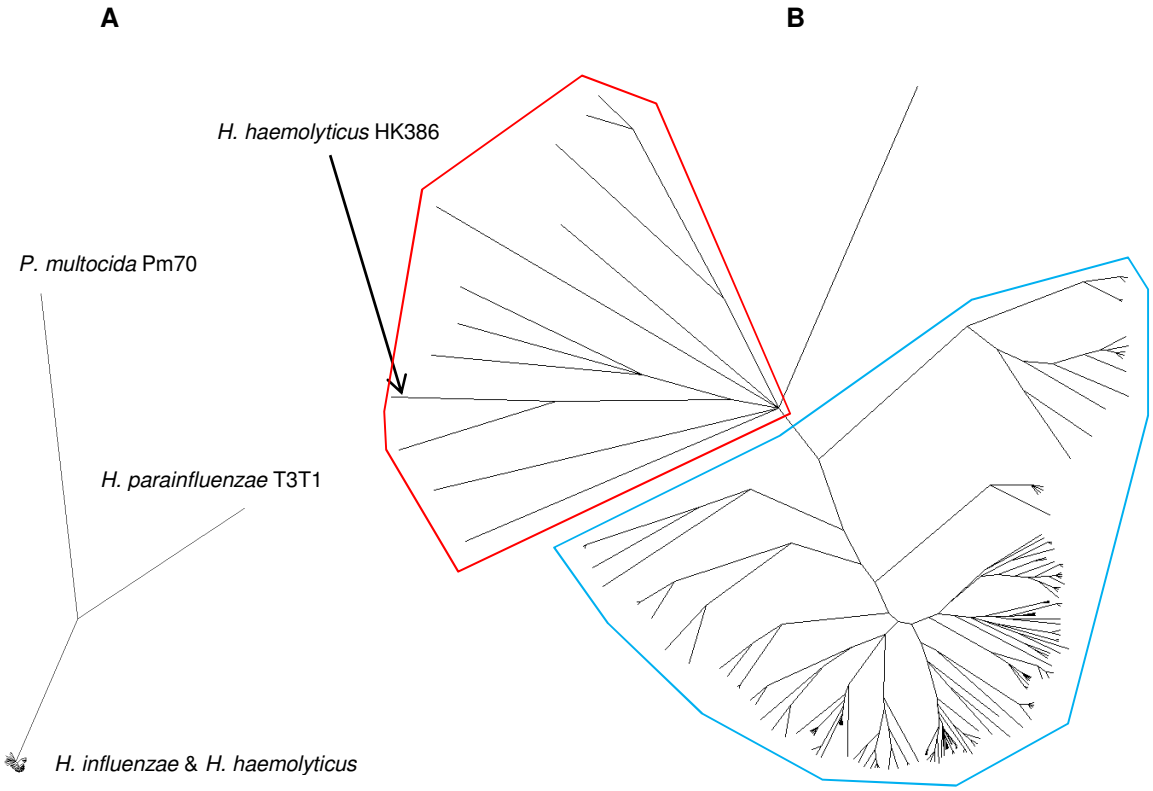
RESULTS

Isolate Characteristics

All isolates were extensively tested using the exclusionary criteria described in the Methods section to ensure that all were 'true' nontypeable *H. influenzae*. Five isolates were missing from the laboratory's bacterial collections and could not be included. Presence of the typeable specific *cap* locus was tested by PCR using the protocols of Davis et al. and Falla et al. (18, 31); eight typeable isolates were discovered and excluded from further analysis. Four isolates with superimposed peaks on sequence chromatograms that were not resolvable by re-isolation of genomic DNA from a single colony of the original bacterial stock were identified and removed from further analysis.

A dataset of six of the seven MLST loci (excluding *fucK*) from 181 of the original 204 selected isolates was analyzed in ClonalFrame. Excluded were the five missing isolates, eight typeable isolates, four persistently contaminated isolates, and six (five US commensal, one US OM) isolates identified as non-Hi (75). In addition, three non-Hi strains from the *Pasteurellaceae* family were included as outgroups: *H. haemolyticus* strain HK386 (98), *H. parainfluenzae*

Figure 3.1



Unrooted majority consensus tree of 181 *Haemophilus* isolates. **A**. Zoomed out view showing the relative positions of the four species. **B**. Zoomed in view illustrating the 11 isolates that cluster with *H. haemolyticus* (circled in red) compared with the remaining 170 NTHi isolates (circled in blue).

strain T3T1 (GenBank ID FQ312002.1), and *Pasteurella multocida* strain Pm70 (74). An unrooted majority consensus tree was constructed by SplitsTree 4.11.3 from the 1,000 trees in the posterior sample (**Figure 3.1**). Eleven of the 181 isolates clustered with *H. haemolyticus* strain HK386, leaving a final total of 170 true NTHi isolates out of the original 204 selected isolates in the final dataset. In total, 34 isolates (seven OM, 27 commensal) were either missing from the collection, typeable, non-*H. influenzae*, or persistently contaminated; these isolates were excluded from further analysis. The above information is

Table 3.1

	Initial[a]	Missing[b]	Typeable	Non-Hi[c]	Contam.[d]	Total Removed	Final	Isolation Date	Age[e]
Finland OM	30	0	0	0	0	0	30	1994-96	2 - 24
Finland Commensal	30	0	4	0	0	4	26	1999	10 - 24
Israel OM	30	0	1	0	1	2	28	2000-01	6 - 48
Israel Commensal	30	0	3	2	3	8	22	2001-02	1 - 59
US OM	42	4	0	1	0	5	37	1996-2001	7 - 84
US Commensal	42	1	0	14	0	15	27	1998-2001	< 36
Total OM	102	4	1	1	1	7	95	1994-2001	2 - 84
Total Commensal	102	1	7	16	3	27	75	1998-2002	1 - 59

Characteristics of isolate collections from Finland, Israel, and the US.
[a] Number of isolates randomly selected from the collection.
[b] Number of selected isolates missing from the collection.
[c] Number of selected isolates designated as non-*H. influenzae*.
[d] Number of selected isolates with unresolvable contamination.
[e] Age range in months of the children from whom the isolates were collected.

enumerated in **Table 3.1**.

One striking aspect obvious from **Table 3.1** is that all but one of the isolates identified as non-*H. influenzae* were in the commensal groups, with the overwhelming majority (14/16) in the US commensal subgroup. This is likely due to differences in study design and priorities. Commensal isolates from Finland (136) and Israel (44) were both collected in health care settings (e.g. hospitals, primary clinics, etc.) with a limited number of isolates obtained from a given child. In general, the investigators conducting these studies were more interested in whether any NTHi was present in the sample, rather than its potential diversity. In contrast, commensal isolates from the US (35, 128) were obtained from healthy children at a number of daycare centers in Michigan, and up to 30 isolates per child were collected. Investigating the diversity of the NTHi present

among those children was one of the goals of the studies. Thus, the US commensal subgroup of isolates may represent a more complete picture of the flora inhabiting the naso- and oropharynges of a healthy child, including examples of what we are now able to distinguish as something closely related to, but separate from, nontypeable *H. influenzae sensu stricto*.

MLST

The 170 NTHi isolates in the final dataset were genotyped by MLST, yielding a total of 109 STs, 53 of which were previously undescribed in the MLST database. Of the 109 total STs, 45 were found only in OM isolates, 51 were found only in commensal isolates, and 13 were found in both OM and commensal isolates. Of the 53 previously undescribed STs, 20 were found only in OM isolates and the remaining 33 were found only in commensal isolates. Fifteen isolates could not be amplified with the *fucK* primers, and were found to be missing the entire fucose operon after following the protocol of Ridderberg et al. (113). Currently, due to technical reasons the MLST database is not able to accept isolates missing a locus for sequence type assignment. Consequently, all isolates missing *fucK* have been assigned a placeholder ST starting at 10,000. General characteristics of the MLST genotyping are detailed in **Table 3.2**, and the specific ST assigned to each isolate is listed in **Table 3.3**.

eBURST Analysis

As in the eBURST analyses conducted in Chapter 2, STs differing from

Table 3.2

	Isolates	*fucK* (-)ve Isolates	*fucK* (-)ve STs	Total STs	New STs
Finland OM	30	0	0	21	3
Finland Commensal	26	0	0	24	6
Finland Total	56	0	0	45	9
Israel OM	28	0	0	22	13
Israel Commensal	22	2	2	21	13
Israel Total	50	2	2	43	26
US OM	37	1	1	25	4
US Commensal	27	12	9	24	14
US Total	64	13	10	49	18
Total OM	95	1	1	45	20
Total Commensal	75	14	11	51	33
Total found in both[a]	NA	NA	0	13	0
Overall Total	170	15	12	109	53

General characteristics of the MLST genotyping.
[a] Listing of STs containing both OM and commensal isolates.

another ST by only one locus are connected by lines and form a group, known as a clonal complex. The most conservative definition of a clonal complex was used, where STs are included in the group only if they share alleles at a minimum of six of the seven loci with at least one other ST in that group. The size of the circles is proportional to the abundance of the corresponding STs in the data set, and it is important to note that the relative placement of unconnected STs is random. The MLST data from the final dataset of 170 NTHi isolates was analyzed with eBURST v3 (**Figure 3.2a**). The result is quite similar to the analyses performed in the previous chapter on isolates from children 22 and 26, as well as all NTHi isolates in the MLST database at that time, in that most STs are not closely enough related to another ST in the sample to form a complex and are thus unconnected. Due to this, there is little clustering evident by either disease (OM/commensal) or geographic region. The largest clonal

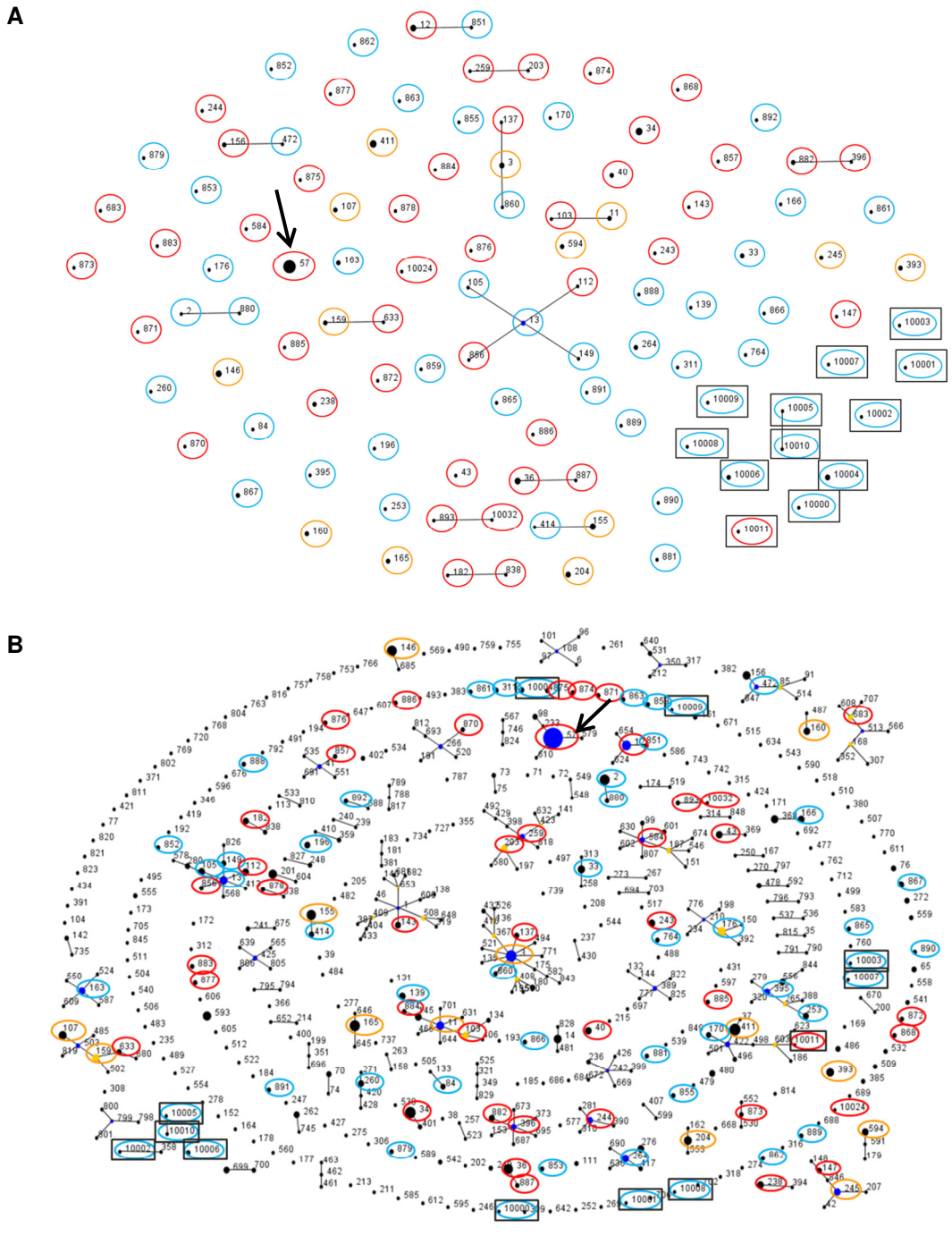
Table 3.3

Finland OM		Finland Commensal		Israel OM		Israel Commensal	
Isolate	ST	Isolate	ST	Isolate	ST	Isolate	ST
F162-7	411	F120	472	I153	165	I242	166
F164-6	3	F202	160	I156	393	I247	311
F199-3	57	F242	33	I162	868	I249	859
F206-4	155	F282	851	I164	57	I255	890
F251-1	245	F285	139	I166	870	I256	764
F286-5	155	F324	13	I167	894	I258	10000
F433-3	40	F441	253	I168	12	I263	860
F567-9	103	F443	411	I169	57	I264	861
F608-5	12	F571	159	I175	871	I270	891
F658-2	156	F638	163	I179	411	I276	862
F885-7	838	F651	852	I181	204	I278	863
F1124-2	12	F894	155	I183	872	I280	892
F1152-8	40	F938	853	I184	873	I283	165
F1232-4	34	F994	411	I185	874	I289	393
F1268-9	36	F1015	245	I187	875	I307	264
F1296-5	57	F1060	260	I188	876	I312	865
F1388-5	36	F1115	411	I191	238	I316	204
F1449-9	36	F1158	163	I198	877	I328	10001
F1541-1	856	F1248	13	I202	57	I336	866
F1588-3	57	F1308	855	I207	895	I338	867
F1618-4	259	F1448	264	I208	893	I340	867
F1663-8	43	F1450	888	I210	878	I345	84
F1702-5	160	F1799	105	I213	244		
F1726-1	147	F1831	395	I218	238		
F1778-9	857	F1942	3	I221	57		
F1897-7	159	F1983	889	I224	57		
F2025-4	57			I226	57		
F2037-2	57			I230	396		
F2188-8	243						
F2206-6	137						

US OM				US Commensal			
Isolate	ST	Isolate	ST	Isolate	ST	Isolate	ST
K7LE2.5	10011	P25RE2.7	57	26.3-23	10002	D07.2.6	33
K8RE2.1	156	P26RE2.2	3	28.4-21	176	E12.2.11	107
K15RE2.6	11	G123	34	30.2-24	2	E14.2.16	10007
K16RE2.4	584	G322	57	37.3-21	146	F05.2.3	10008
K17ME2.5	146	G423	34	39.4-23	10003	G06.2.1	594
K19RE2.4	57	G522	36	54.3-22	155	H04.2.1	11
K21LE2.7	882	G622	884	55.2-22	10004	J06.2.2	880
K26LE2.7	683	G723	885	59.3-25	10005	M01.2.5	10009
K27RE2.8	146	G822	57	61.4-22	10004	M02.2.4	414
K29RE2.10	112	G922	203	62.2-24	10004	N02.2.3	881
K32RE2.5	143	G1023	886	63.4-22	196	O07.2.12	149
K33RE2.4	146	G1123	34	65.2-23	170	P11.2.3	10006
K34LE2.1	882	G1222	887	C09.2.3	10006	P20.2.1	10010
K35LE2.1	883	G1322	633	C17.2.11	879		
P5ME2.5	204	G1423	594				
P6RE2.9	182	G1522	57				
P16LE2.7	12	G1623	57				
P19LE2.6	57	G1822	34				
P20LE2.10	107						

MLST sequence type assignments. Isolates missing the fucose operon have placeholder ST numbers 10,000 and above.

Figure 3.2



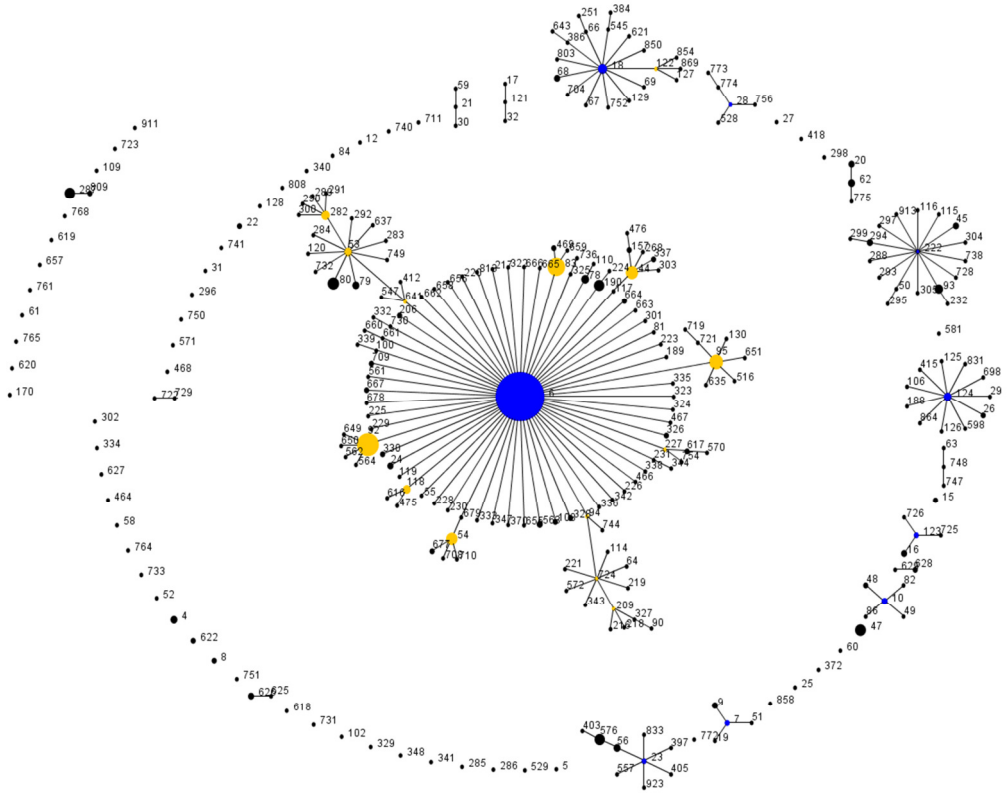
eBURST analyses of NTHi isolates. The key at the bottom identifies aspects of the STs identified in this study. The thick black arrows point to ST57. **A**. Analysis of the 109 STs found in the 170 NTHi isolates of the final dataset. **B**. Analysis of all 537 NTHi STs found in the MLST database (accessed 03-31-11) and the 12 *fucK* negative STs from part A. The largest clonal complex consists of 19 STs.

complex consists of only five STs, and contains both OM and commensal STs. Only ten STs contain three or more isolates, and in eight of these STs the isolates are from at least two of the geographic regions. By far the most common ST is ST57, which represents 18 of the 170 NTHi isolates in the final dataset, and 28 of all the NTHi isolates in the MLST database. Intriguingly, all 18 ST57 isolates in the final dataset were collected from cases of otitis media, and were gathered in approximately equal numbers from each of the three geographic regions (Finland: 5 isolates; Israel: 6 isolates; US: 7 isolates).

This high level of diversity remains consistent when all 537 NTHi STs (836 isolates) in the MLST database circa March 31st, 2011, as well as the 12 *fucK* negative STs (15 isolates), were analyzed (**Figure 3.2b**). The 537 STs include all 97 *fucK* positive STs identified in this study. Again, little clustering is evident, and the few complexes are rather small (the largest being composed of 19 STs). This suggests that the high diversity observed in the collection of 170 isolates from three geographic areas is not merely an artifact of incomplete sampling, but may instead represent the true diversity of NTHi.

This high diversity can be contrasted with an eBURST plot of all typeable STs in the MLST database as of June 28th, 2011 (**Figure 3.3**). Here, the picture is quite different, with nearly half (126, or 45%) of the 281 STs comprising a single clonal complex, with many of the rest forming smaller groups. It should be noted that type b STs predominate, making up 2/3 of all typeable STs in the database (187 of 281). This includes the central clonal complex observed in Figure 3.3, where all 126 STs are type b.

Figure 3.3



eBURST analysis of all 281 typeable STs in the MLST database (accessed 06-28-11). The central clonal complex consists of 126 type b STs.

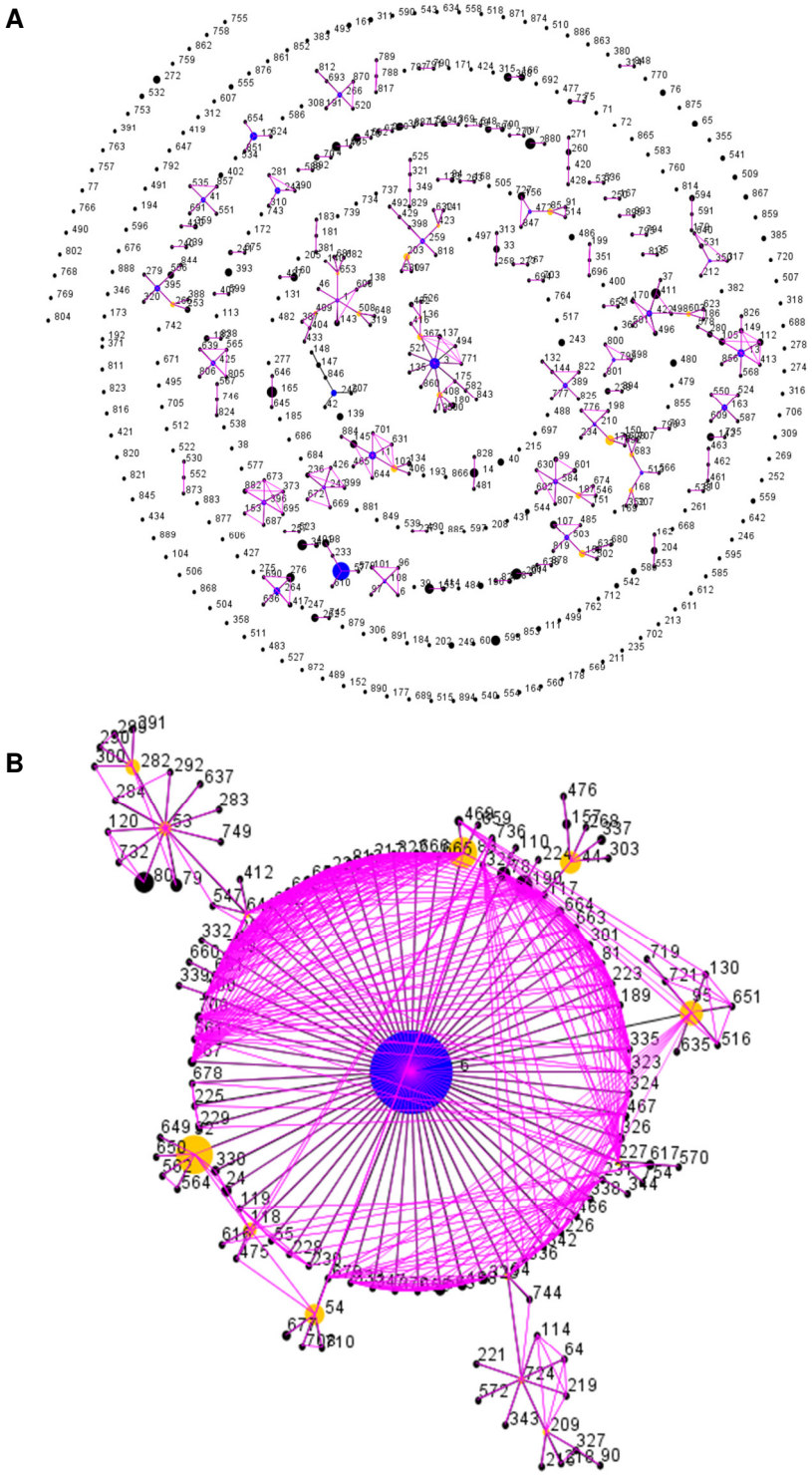
Turner et al. have examined eBURST performance in a range of simulated and real datasets, and have found three patterns that may indicate poor reliability of the clustering (135): 1) a single large, straggly group; 2) long range undrawn SLV links across the group; and 3) a high proportion of STs in the largest group (greater than 25%). In many cases, these patterns are caused by a high rate of recombination relative to mutation. In their analysis of MLST data from 18 bacterial species, these criteria enabled them to identify five species in which eBURST performance is likely to be poor.

It is obvious from **Figure 3.2** that a single large, straggly group does not

predominate within NTHi. As discussed earlier, there is hardly any grouping at all. One large group does dominate the eBURST plot of typeable STs, but it is radial rather than straggly and quite different from those discussed by Turner et al. **Figure 3.4a** shows an eBURST analysis of all NTHi STs in the MLST database (accessed June 28th, 2011) with all SLV links drawn in pink, including those normally undrawn by the program. **Figure 3.4b** shows a similar plot for the Hib central clonal complex from **Figure 3.3**. In both figures, the undrawn SLV links do not appear to be of the type suggested by Turner et al. to indicate unreliability.

Turner and colleagues calculated the proportion of STs in the largest eBURST group for *H. influenzae* at approximately 15%, which put it in the range where eBURST performance should be optimal (135). However, they committed an error common to those who are not familiar with Hi by combining typeable and nontypeables for their analysis. The differences in between the populations of typeable and nontypeable Hi have been discussed earlier in this work, and are readily apparent simply by examining eBURST plots and MLST statistics. For example, despite having similar numbers of isolates (836 NTHi, 718 typeable) in the MLST database, there are nearly twice as many NTHi STs (537 NTHi, 281 typeable). It would have been more correct to separate them in assessing the performance of eBURST. For NTHi, the largest group among the 537 STs in **Figure 3.2b** is 19, giving a proportion of 3.5%. This places NTHi with *H. pylori* as examples of bacteria in which the population is so diverse that clonal complexes are not apparent, likely due to high rates of both recombination and mutation

Figure 3.4



eBURST analyses showing all possible SLV links in pink. **A**. All NTHi STs in the MLST database as of 06-28-11. **B**. The Hib central clonal complex from Figure 3.3.

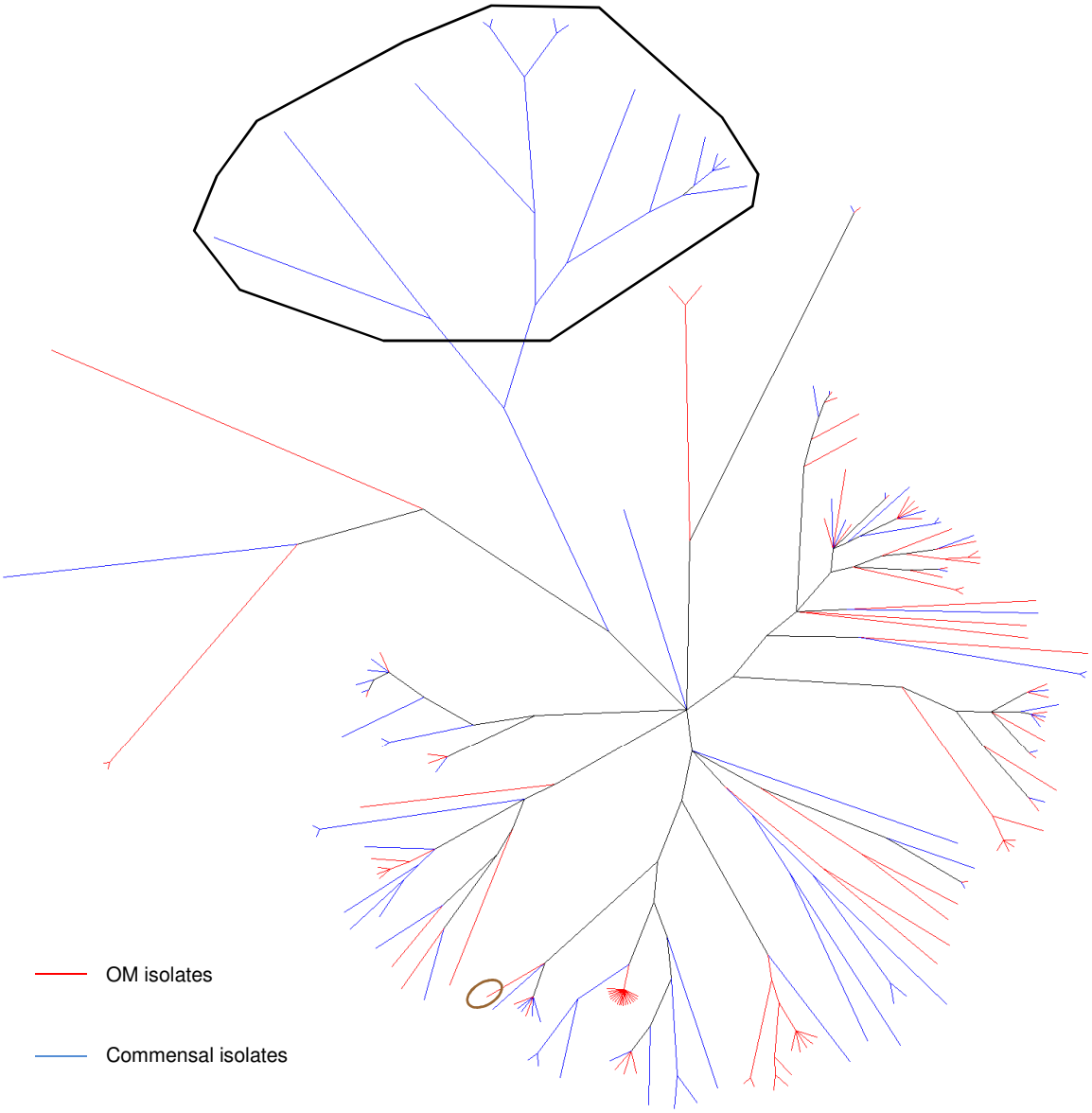
(135). Conversely, the largest group among the 281 typeable STs in **Figure 3.3** is 126, yielding a proportion of 44.8% and putting them near the range with potentially questionable performance (though it is still likely to be acceptable as the main group is radial rather than straggly, and typeable isolates typically have lower rates of recombination (102)).

Phylogenetic analysis

ClonalFrame was used to infer the clonal relationships between the 170 NTHi isolates in the final dataset. All seven MLST loci were utilized, with isolates missing the *fucK* gene treated as having a long gap at that locus. SplitsTree 4.11.3 was used to create the unrooted majority consensus tree from the 2,000 trees in the posterior sample. **Figure 3.5** shows this tree color coded by disease status (OM/commensal), while **Figure 3.6** show the same tree color coded by geographic region (Finland/Israel/US). As in the analyses with eBURST, only very limited clustering is apparent by either disease status or geographic region. The majority of clades are composed of both OM and commensal isolates from multiple geographic regions, suggesting that neither of these factors imposes an overwhelming constraint on the isolates in this sample.

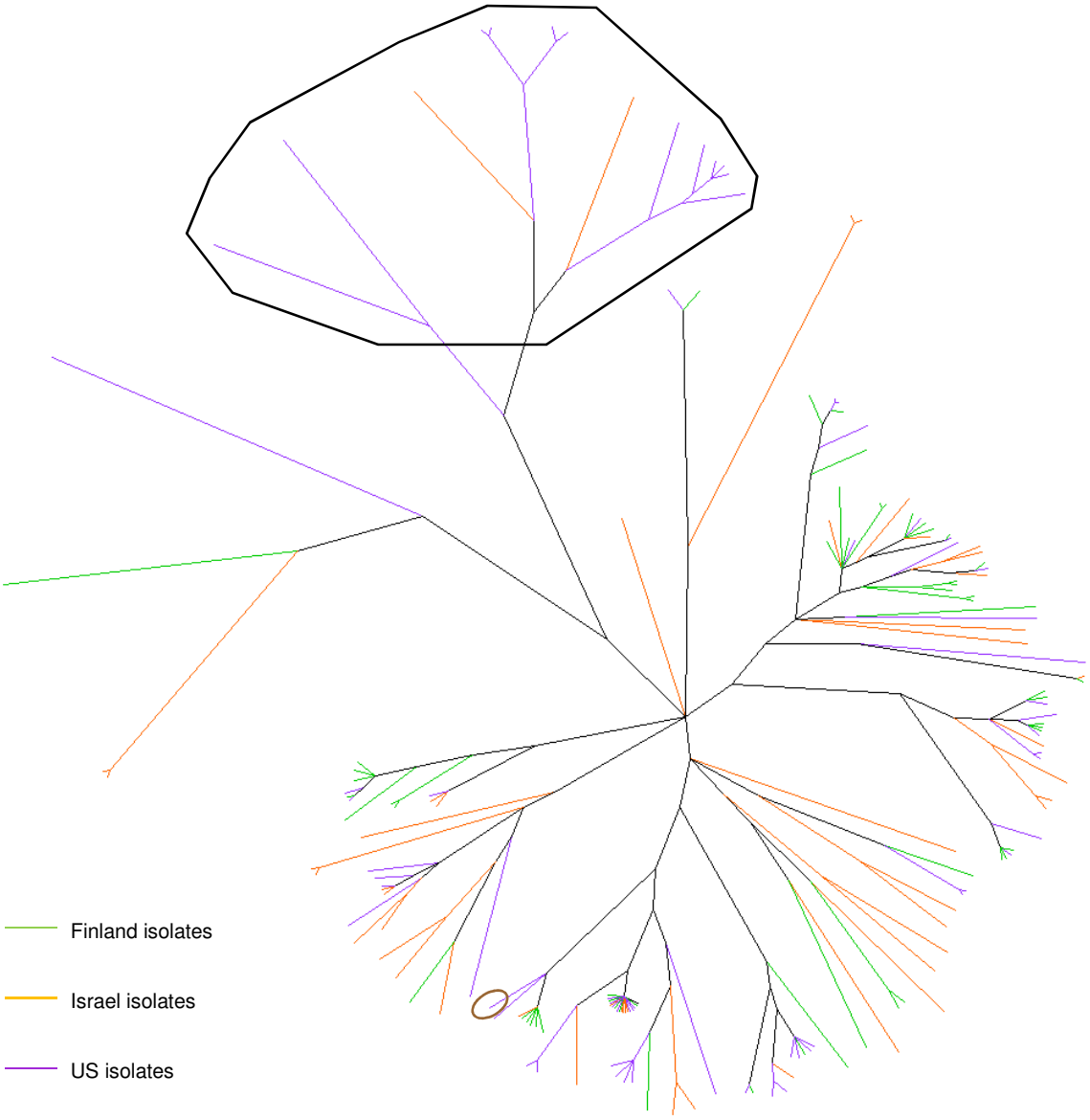
One exception to this is the commensal *fucK* negative group of isolates, circled in black in **Figures 3.5** & **3.6**. In contrast to the OM *fucK* negative isolate (circled in brown), which falls in the midst of other NTHi isolates in the core of the tree, the commensal isolates form a distinct cluster that is fairly distant from the majority of the tree. This group can also be seen in **Figure 3.1b** on the far

Figure 3.5



Unrooted majority rule consensus tree constructed from the MLST data from 170 NTHi isolates, colored by disease. Branches connecting only to OM isolates are colored red, while branches connecting only to commensal isolates are colored blue. Commensal *fucK* negative isolates are circled in black, while the OM *fucK* isolate is circled in brown.

Figure 3.6



Unrooted majority rule consensus tree constructed from the MLST data from 170 NTHi isolates, colored by geographic area. Branches connecting only to Finland isolates are colored green, branches connecting only to Israel isolates are colored orange, and branches connecting only to US isolates are colored purple. Commensal *fucK* negative isolates are circled in black, while the OM *fucK* isolate is circled in brown.

middle right of the tree, where they occupy a position between *H. haemolyticus* and the rest of the NTHi. This supports the increasingly popular idea of a continuum between closely related bacterial species, rather than tidy distinct definitions that strains will neatly occupy (126). They seem to deserve classification as NTHi, however, because, apart from the phylogenies presented here, six of the 14 isolates have been tested for the presence of the *iga* and *lgtC* loci by microarray, and all six were positive (data not shown). Reactivity against probes for these genes has been shown to be a robust discriminatory marker for distinguishing *H. influenzae* from *H. haemolyticus* (75), and is part of our laboratory's definition for what is accepted to be true NTHi.

ClonalFrame was also used to investigate the relative rates and contributions of recombination and mutation in the sample in the form of two ratios: ρ/θ and r/m. ρ/θ is the ratio of the recombination rate to the mutation rate, and is therefore a measure of how often recombination events occur relative to mutations. However, as a single recombination event could potentially introduce many more nucleotide changes than a mutation, the rate at which each of the two processes occurs is perhaps not the most informative measure. More interesting perhaps would be an indicator of how important recombination and mutation are in the evolution of the sample. The ratio of probabilities that a given site is altered via recombination or mutation, r/m, is such a statistic. In effect, it is a direct measure of the importance of recombination relative to mutation in the diversification of the sample. These measures for this sample, as well as δ (the average tract length of a recombination event), are reported in **Table 3.4**.

Table 3.4

ρ/θ[a]	r/m[b]	δ[c]
1.00	5.05	493.43
0.67 - 1.44[d]	3.62 - 6.93[d]	383.05 - 633.16[d]

Ratios of the rate and effect of recombination versus mutation inferred by ClonalFrame.
[a] ratio of the rates of recombination versus mutation
[b] ratio of the effects of recombination versus mutation
[c] average tract length of a recombination event
[d] 95% credibility regions

The ratio of the rates of recombination and mutation (ρ/θ) is one, indicating that the two processes occur at approximately the same rate. However, the ratio of the probabilities that a given nucleotide is changed by recombination or mutation (r/m) is 5.05. This indicates that despite both processes occurring at the same rate, recombination introduces over five times more nucleotide substitutions than do point mutations. The higher rate and impact of recombination than mutation in this sample of NTHi is consistent with data reported in Chapter 2 and by others (17, 102, 141).

Population Structure

The *structure* 2.3.3 program was used to assess population structure in the sample. Two datasets were utilized, one including only a single example of each of the 109 unique STs found previously (unique STs dataset), and the other including all 170 NTHi isolates (all isolates dataset). For both datasets, 20 independent runs were performed, and the CLUMPP program was used to assess multimodality among the runs. This is an essential step during the analysis and interpretation of *structure* results, as the algorithm it implements can identify non-symmetric modes (or clustering solutions with high posterior

probabilities), particularly in complex datasets with large values of *K*. The current implementation of *structure* typically does not cross between these modes, which can lead to different runs producing very different answers (110). Within each distinct mode for a given *K*, the estimated log probability of observing the data (lnP(D)) and individual membership proportions (*Q*) were averaged.

Multimodality was indeed apparent within the two datasets for many values of *K*, from a high of 12 distinct modes found among the 20 runs at *K*=6 in the all isolates dataset to a low of a single mode at *K*=2 in the unique STs dataset. A summary of this information is presented in **Table 3.5**. The mode that maximized the lnP(D) at each value of *K* was chosen for further analysis.

Table 3.5

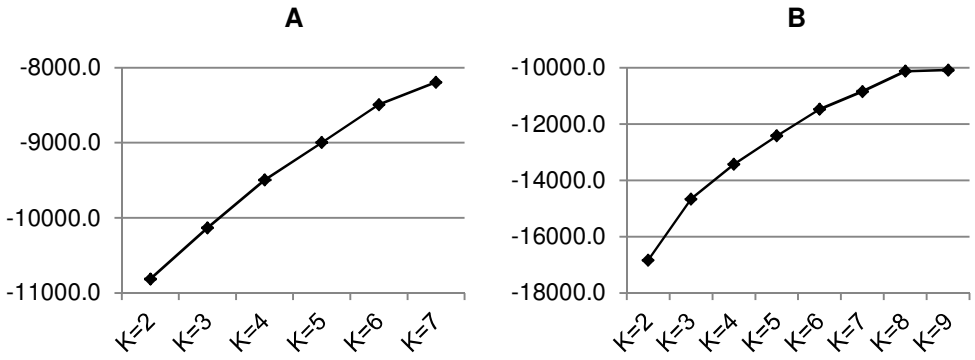
	Unique STs Dataset			All Isolates Dataset		
K	Modes[a]	Max Mean lnP(D)[b]	n[c]	Modes[a]	Max Mean lnP(D)[b]	n[c]
2	1	-10815.3	20	3	-16838.5	9
3	2	-10133.1	10	3	-14668.0	17
4	5	-9496.3	2	6	-13431.6	10
5	5	-8997.6	15	9	-12424.6	8
6	4	-8492.0	1	12	-11479.5	2
7	7	-8194.9	1	11	-10846.6	6
8				4	-10127.2	5
9				4	-10083.9	14

Multimodality in the population structure analysis for both datasets.
[a] Total number of modes identified for that *K*
[b] Average lnP(D) for the mode that maximized this value
[c] Number of replicate runs in the mode with the maximum lnP(D)

Estimating the correct number of populations *K* is not necessarily a straightforward procedure. The *structure* program provides an *ad hoc* method based on computing the posterior probabilities of *K* from the lnP(D) values at different *K*'s. In essence, the *K* at which the lnP(D) (and thus the posterior

probability) plateaus and ceases to increase is considered to be an optimum choice for the number of populations. However, while this method generally works well in datasets with small numbers of discrete populations, it is less helpful in more complex situations. In these scenarios, the lnP(D) often continues to increase with increasing *K* and never plateaus, even well past any biologically meaningful values for the number of populations (110). This continual increasing of the posterior probabilities can be observed for both the unique STs dataset (**Figure 3.7a**) and the all isolates dataset (**Figure 3.7b**), though the plot for the all isolates dataset may be plateauing at $K = 8$.

Figure 3.7



Plots of the maximum lnP(D) versus *K*. **A**. Unique STs dataset. **B**. All isolates dataset

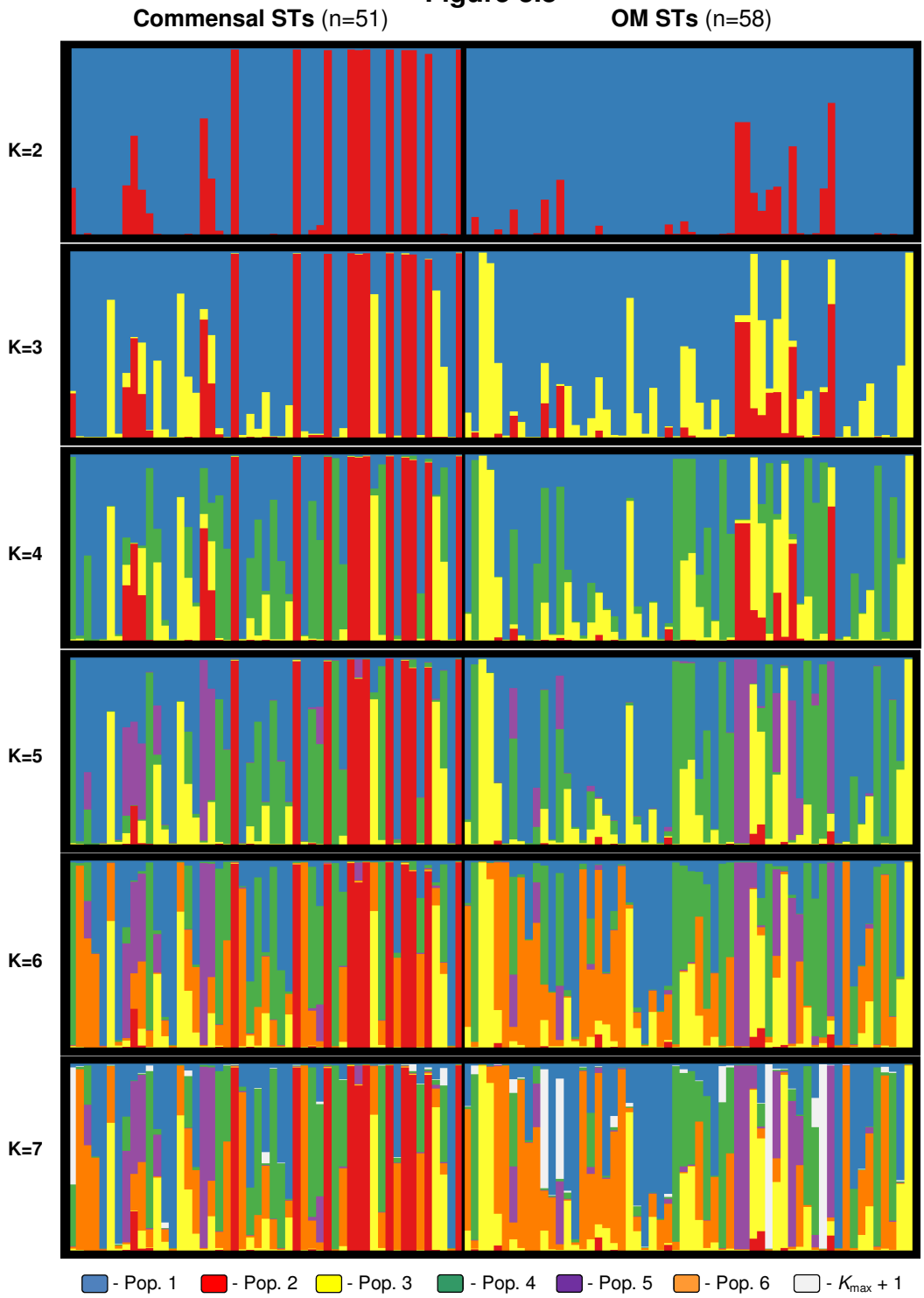
To determine the most appropriate number of populations, a method similar to that used by Verdu et al. was utilized (138). The number of assumed populations was increased and those data assessed until adding another population became uninformative, identified by the individual membership proportions *Q* for that population being on average very low and few individuals having a large portion of their ancestry from that population. For the unique STs

dataset, this occurred at *K* = 7, where the average *Q* was 4.5% and only two STs had greater than 65% of their ancestry from that population. In the all isolates dataset, up to eight populations were well supported; if *K* was increased to nine, the average *Q* for the added population was 2.3% and only two isolates had more than 65% of their ancestry from that population.

Figure 3.8 illustrates the clustering solutions inferred by *structure* from the unique STs dataset that maximize lnP(D) for each value of *K* from 2 to 7. Commensal STs were those STs containing only isolates collected from healthy children in this sample (n = 51), while OM STs contained at least one isolate collected from a case of otitis media (n = 58). The lack of support for adding a population past *K* = 6 can be seen in the bottom panel, where almost none of the of the STs trace their ancestry back to the added population (seen in white) and the clustering solution is otherwise nearly identical to that at *K* = 6. **Figure 3.9** displays the same information for the all isolates dataset, with *K* ranging from 2 to 9. Once again, the lack of support for additional populations past *K* = 8 can be seen in the bottom panel, where extremely few isolates of the sample have membership in the *K* = 9 population (seen in white). The clustering solution at *K* = 9 does differ from the *K* = 8 solution in that far fewer of the isolates have significant membership in the *K* = 8 population (colored green). However, the lack of support for *K* = 9 as well as the apparent plateau in the lnP(D) values at *K* = 8 and 9 indicates that *K* = 8 is the best choice for this sample.

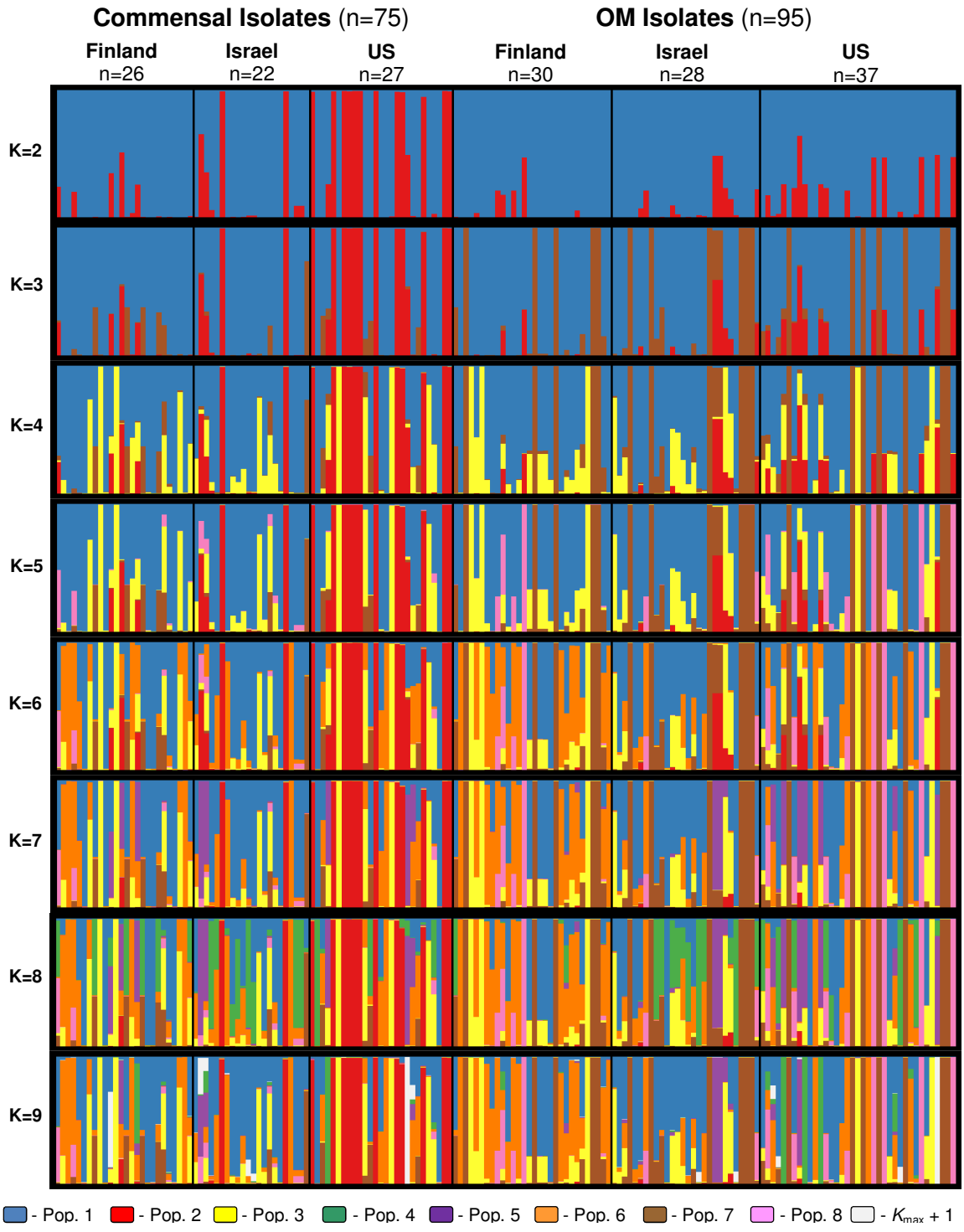
For both analyses, choosing these values for *K* holds to the guidelines stated by the author of the *structure* program, in that we might not know the **true**

Figure 3.8



Population structure inferred by *structure* for the 109 unique commensal and OM NTHi sequence types. The number of predefined populations (*K*) is indicated to the left of each plot. Each ST is represented by a vertical line partitioned into *K* colored components according to the estimated individual membership proportion in each population (*Q*). The average of all replicate runs within the mode with the highest likelihood at each *K* is shown. Population color coding matches Figure 3.9. K_{max} + 1 refers to the first *K* that is no longer informative (i.e. *K*=7).

Figure 3.9



Population structure inferred by *structure* for all 170 commensal and OM NTHi isolates. The number of predefined populations (*K*) is indicated to the left of each plot. Each isolate is represented by a vertical line partitioned into *K* colored components according to the estimated individual membership proportion in each population (*Q*). The average of all replicate runs within the mode with the highest likelihood at each *K* is shown. Population color coding matches Figure 3.8. K_{max} + 1 refers to the first *K* that is no longer informative (i.e. *K*=9).

value of *K*, but we should target the smallest value that captures the major structure in the data (110). This is reinforced by the considerable amount of information gained at $K = 6$ for the unique STs dataset and at $K = 8$ for the all isolates dataset, where the added populations (orange in **Figure 3.8**, green in **Figure 3.9**) comprise a significant portion of the sample's ancestry.

With two exceptions (populations 7 and 8 from the all isolates dataset), the *structure* analyses inferred the same populations for both datasets. When considering those individuals (whether STs or isolates) with a large proportion of their ancestry from one population, they are clustered together in both analyses, though they are not necessarily inferred in the same order. For example, consider two sequence types with greater than 99% of their ancestry from population 6 (orange in **Figures 3.8 & 3.9**) when analyzing the unique STs dataset, ST3 and ST33. ST3 consists of isolates F1942 (Finland commensal), F164-6 (Finland OM), and P26RE2.2 (US OM), while ST33 consists of isolates F242 (Finland commensal) and D07.2.6 (US commensal). In the analysis of the all isolates dataset, these isolates also have greater than 99% of their ancestry from the same population. This trend holds for the vast majority of STs and isolates with a large percentage of their ancestry from one population. There are some differences in assignment of STs and isolates with a more admixed heritage, but overall the similarity of clustering between the two analyses is very high. Henceforth, the populations inferred by *structure* will be referred to by the labels presented in **Figures 3.8 & 3.9**. Both figures are color coded identically, so the red colored population 2 (for example) refers to the same genetic cluster

in both datasets.

One major difference is that analysis of the all isolates dataset identified two additional populations, labeled population 7 (brown) and population 8 (pink) in **Figure 3.9**. Population 7 is comprised of 18 OM isolates of ST57, while population 8 consists of five OM isolates of ST34. This may be a situation similar to those mentioned by the authors of *structure*, in which having multiple family members (or in this case, multiple isolates with the same ST) can lead to an overestimation of *K*, though, as mentioned, there tends to be little effect of the assignment of individuals to populations for a given *K* (33, 110). Indeed, as mentioned above, with the exception of populations 7 and 8, the overall clustering solutions between the two analyses are nearly identical. However, these two clusters, though perhaps not true populations in the usual sense, are of interest as large groups of identical genotypes associated solely with otitis media, despite being found in different times, places, and people.

The complexity of the *structure* plots in **Figures 3.8 & 3.9** at the chosen levels of *K* (six and eight, respectively) make discerning trends in population structure by either geographic area or disease difficult. The most readily apparent features are isolates and STs with a very high proportion of their ancestry from population 2 (in red), which seem to be found exclusively among commensal STs in **Figure 3.8**, and with two exceptions only among US commensal isolates in **Figure 3.9**. Population 2 corresponds exactly to the 14 *fucK* negative isolates (11 STs) identified during MLST genotyping. As *fucK* was not used for the analysis of population structure, their identification as a distinct

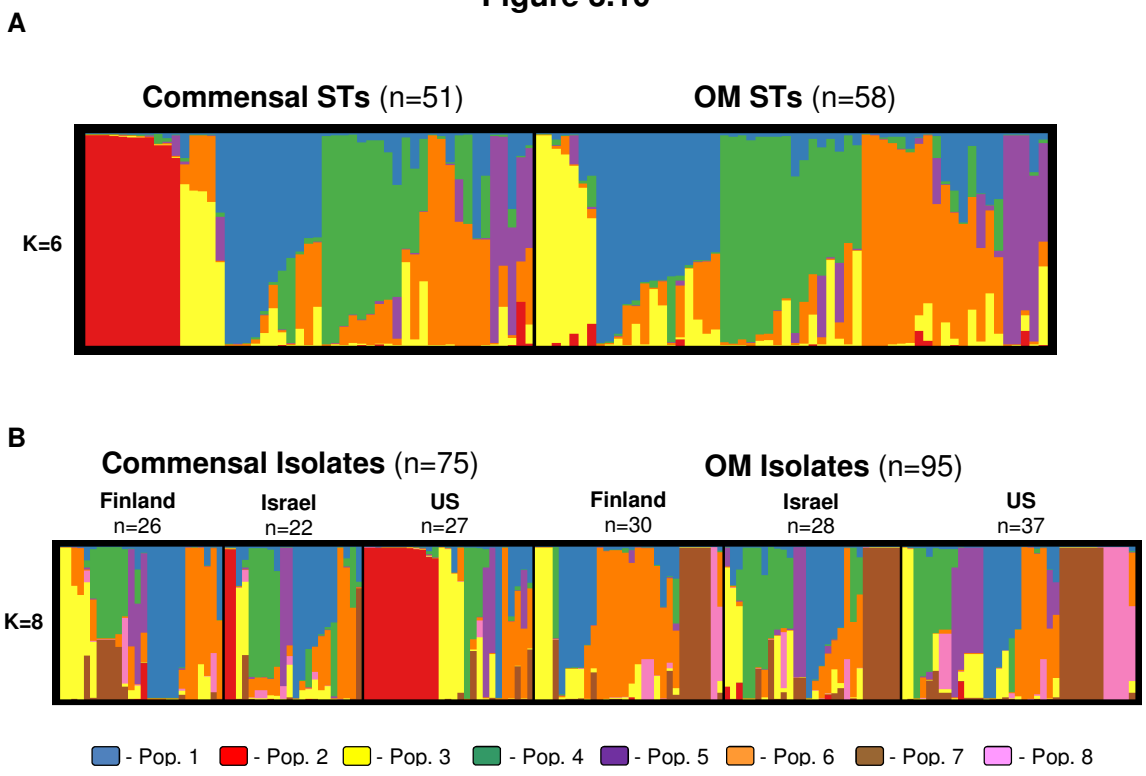
cluster is not biased due to their deletion at that locus. The phylogenetic analysis presented in **Figures 3.5 & 3.6**, which placed these isolates in a distinct clade apart from the remaining NTHi, reinforces this clustering solution.

Rearranging the population structure plots helps to present a clearer picture, as shown in **Figure 3.10**. Panel A presents the $K = 6$ plot from **Figure 3.8** while panel B presents the $K = 8$ plot from **Figure 3.9**; both plots have been sorted by individual membership proportion in each K. From this figure, the clustering of population 2 only among commensal STs and isolates is even more obvious. However, most other populations are fairly evenly distributed between disease states and geographic areas. One exception may be population 6 (in orange), which appears to be larger among OM STs in this sample. In terms of STs with a high proportion of their ancestry from that population, both commensal and OM groups have five STs with a Q greater than 75%, but the OM group has an additional eight STs with a Q greater than 60% compared to only two from the commensal group. Panel B offers a more nuanced picture, showing that the greater abundance of population 6 STs in the OM group can be traced to the Finland OM group, which has seven isolates with a Q greater than 75%. The groups with the next highest number of isolates with a population 6 Q greater than 75%, the Finland commensal and US OM collections, have only three apiece.

The populations inferred by *structure* can also be mapped onto the phylogeny estimated by ClonalFrame, as shown in **Figure 3.11**. Seven of the eight populations identified in the all isolates dataset correspond to monophyletic

groups, indicating that both programs arrived at similar conclusions regarding the ancestry of those isolates. However, population 5 (in purple) is polyphyletic and mapped onto portions of three separate clades. This could denote a greater uncertainty or difficulty in estimating the ancestry of these isolates. Alternatively, it could simply be an illustration of differing results from using the different methods implemented by the two programs.

Figure 3.10



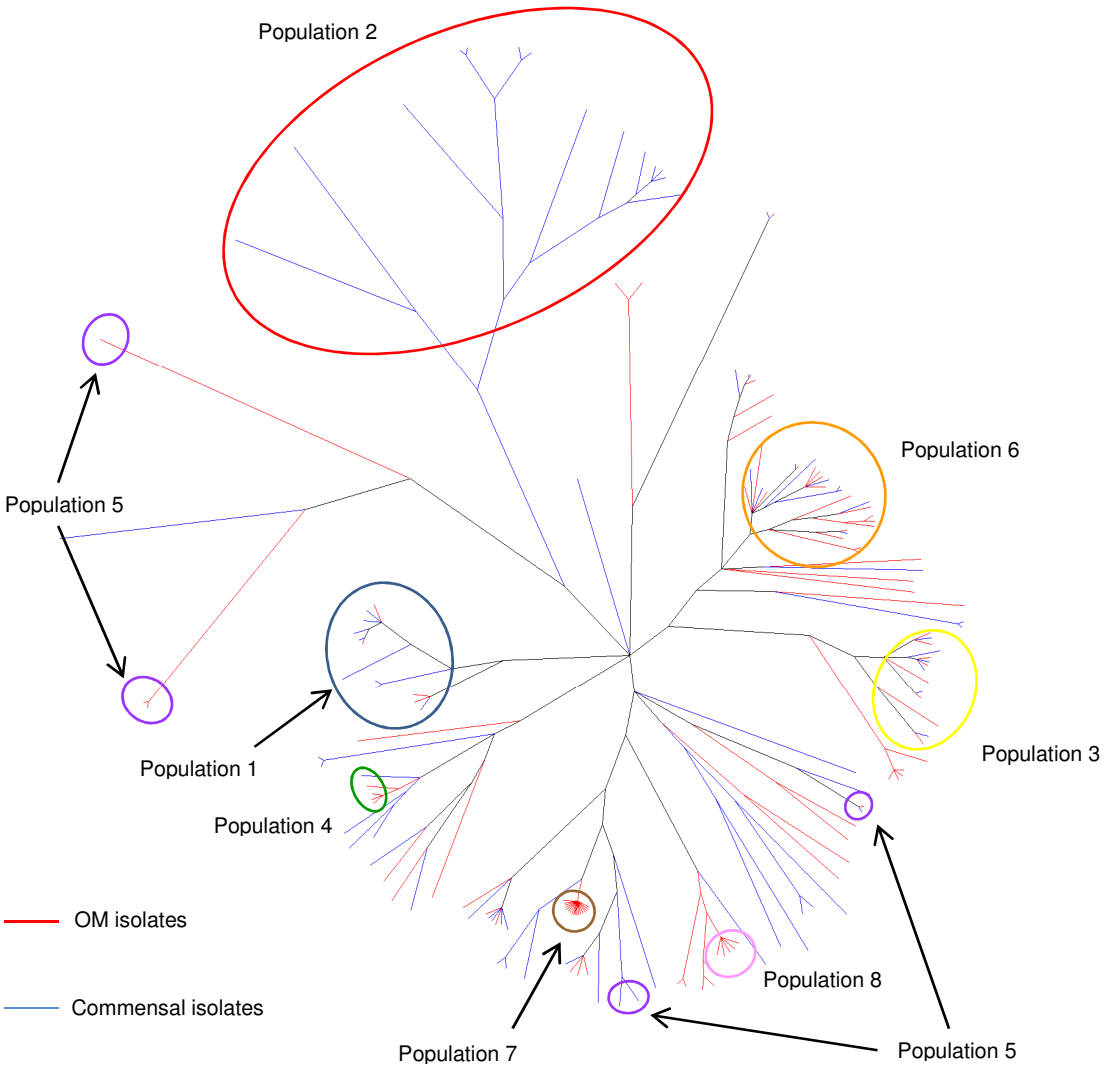
Inferred population structure from Figures 3.8 and 3.9, sorted by individual membership proportion. The number of predefined populations (*K*) is indicated to the left of each plot. Color coding is consistent between panels A and B. **A**. Unique STs dataset. **B**. All isolates dataset.

DISCUSSION

Advances in molecular biology and bioinformatics have greatly aided the investigation of bacterial population structure, leading to a number of interesting publications, including the fascinating report from Falush et al. described in Chapter 1 that was able to infer ancestral populations of *H. pylori* whose spread can be mapped to historical human migrations (34). Recently, Sheppard and colleagues expanded on their work characterizing the convergence of *Campylobacter jejuni* and *C. coli* (125) and found that despite the high diversity of the two species, strong structuring of the populations by host source was apparent, which was stronger than the structuring by geography (124). Budroni et al. investigated 20 full genome sequences for *N. meningitidis* and found evidence for a population structured into phylogenetic clades, despite high rates of detectable recombination throughout the bacterial genome (11). Intriguingly, they identified 22 restriction modification systems whose distribution coincided with the clades, suggesting that the observed population structure may in part be due to a differential barrier to gene flow generated by the restriction systems.

The population structure of nontypeable *H. influenzae* has been less well characterized. As described in the introduction to this chapter, much of the research in this field relied upon MLEE and concentrated on typeable isolates, including only a nominal collection of NTHi at best. However, they observed that while the population of typeable Hi, and type b in particular, is clonal, the population of NTHi is large, heterogeneous, and distinct from that of type b isolates (91-95, 104). More recently, PFGE has been used to again find a clonal

Figure 3.11



Consensus tree from Figure 3.5 with populations inferred by *structure* circled. Isolates with a high proportion of their ancestry from a given population were considered to be members of that population and thus included in that circle. Population colors are coded as in Figures 3.8 – 3.10.

population structure in Hib isolates from Australian Aborigines and non-Aborigines (78) and to identify a lack of change in the population of Hib genotypes causing vaccine failures in the United Kingdom as compared to genotypes found in the pre-vaccine era (3). The study by Erwin and colleagues identifying well defined phylogenetic groups of NTHi based on MLST data has been discussed in the introduction (27).

The research presented here investigated the diversity and population structure of a collection of both commensal and otitis media-associated NTHi. MLST was used to genotype 170 NTHi isolates from Finland, Israel, and the US, identifying 109 unique STs, of which 53 were previously undescribed. Like the study by Musser et al. (91) using MLEE to compare NTHi to Hib isolates (and unlike the similar study by Porras et al. (104)), no ST found among the NTHi isolates was shared by any of the eight identified typeable isolates. This discrepancy has several potential causes, such as a lack of discriminatory power caused by their use of only a small number of MLEE loci (six versus the 15 used by Musser) or their potential misidentification of typeable strains as nontypeable, including capsule deficient strains of a typeable background (18). The majority of STs (81 of 109) were found only once, and only ten of the remaining STs were found three or more times. Thirteen STs were comprised of both commensal and OM isolates, which is not unexpected. The naso- and oropharynges are the reservoir of NTHi, from which emerge strains able to travel up the Eustachian tubes to the middle ear space and initiate otitis media. Thus, one would anticipate isolating genotypes otherwise implicated in disease from the naso- and

oropharynges of currently healthy children.

Interestingly, 15 isolates comprising 12 STs were found to have a deletion of the entire fucose operon, which includes the MLST locus *fucK*. This phenomenon has been previously described in two isolates by Ridderberg et al., who found that one isolate clustered with confirmed Hi in their phylogenetic tree, while the other isolate occupied a more intermediary position between Hi strains and the so-called 'variant' strains (including non-hemolytic *H. haemolyticus*) (113). The phylogeny constructed in ClonalFrame (**Figures 3.5 & 3.6**) presents a similar picture, with the single OM *fucK* negative isolate falling in the core of the tree and the remaining 14 commensal *fucK* negative isolates forming a more distant clade. In **Figure 3.1b**, this group, while still among the NTHi portion of the tree, can be seen to occupy an intermediate position between NTHi and the closely related *H. haemolyticus*. This supports the position that distinct divisions between closely related bacterial species frequently offer an inadequate summation of the true relationships among the bacteria (126).

Analysis with the eBURST revealed that few STs were related closely enough to form clonal complexes, with no significant clustering by either disease or geographic area. To be a member of a clonal complex, a ST had to be identical to at least one other ST in the complex at a minimum of six of the seven MLST loci. The largest group consisted of only five STs, and contained both commensal and OM isolates from multiple geographic areas. Analysis of all 537 NTHi STs in the MLST database confirmed this high level of diversity, with the largest group composed of only 19 STs. These findings are similar to those

presented in Chapter 2, and show that the considerable expansion of the database in the years since that analysis and those presented by Erwin et al. (27) have not provided reason to alter our understanding of NTHi as a very diverse organism. Further credence is provided by the more consistent sampling scheme employed in this study compared to the MLST database as a whole. While the 170 isolates came from multiple studies with various original goals, care was taken to match the isolates, in an approximate manner, on time, space, and age of the host children, and only one isolate was selected per child. This last criterion is an important point, as it means that while multiple isolates with the same MLST genotype were identified, they do not constitute clones in the traditional molecular biologic sense, as they were collected from different individuals at different times, and often from very different parts of the world.

The phylogenetic analysis performed in ClonalFrame further demonstrates the high diversity of NTHi. Some distinct clades were identified, but similar to the results reported by Erwin et al. (27) and those apparent from the eBURST analyses, the clustering does not seem to be predicated on either site of isolation or geographic location. With the exception of the commensal *fucK* negative isolates, the clades are composed of both commensal and OM isolates from multiple geographic regions.

ClonalFrame was also used to estimate the relative rates and contributions of recombination and mutation. While the relative rates of recombination and mutation were approximately equal, recombination was estimated to introduce over five times more nucleotide substitutions than

mutation (r/m = 5.05). This ratio is higher than that found by Vos and Didelot using the same method (r/m = 3.7) (141). However, similar to the previous discussion on eBURST, the authors combined both typeable and nontypeable isolates in their analysis. Perez-Losada et al. found that NTHi have higher rates of recombination at four of the seven MLST loci than do typeable isolates (102), which may account for the lower r/m ratio found by Vos and Didelot. Perez-Losada et al. also reported a measure of recombination relative to mutation (the per allele ratio of recombination to mutation Γ/Γ_{Wf} = 2.51), but as they utilized a completely different analytic technique, comparing values between the studies is difficult. The higher influence of recombination on the evolution NTHi relative to mutation supports the theory that recombination features prominently in the evolution of the species.

The population structure of the sample was further assessed using the *structure* program. Two datasets were analyzed, one with all 170 isolates and one with only a single example of each unique ST. Multimodality was observed at most levels of *K* when analyzing both datasets, and the mode that maximized the log probability of the data at each *K* was used. The smallest number of populations *K* that captured the major population structure within the data was chosen for each dataset (*K* = 6 for the unique STs dataset, *K* = 8 for the all isolates dataset). With two exceptions discussed below, the essentially the same populations were inferred in both datasets. Thus, an isolate with the majority of its ancestry from population 1 based on analysis of the all isolates dataset, for example, will belong to ST that has the majority of its ancestry from population 1

in the analysis of the unique STs dataset. This demonstrates that the inclusion of multiple genetically identical bacterial isolates in the all isolates dataset did not substantially alter the clustering solutions inferred by *structure* and that the information provided by both analyses is largely identical.

Despite the high genetic diversity of NTHi described throughout this thesis, significant population structure was apparent in this sample. As can be seen in **Figures 3.8 – 3.10**, there are numerous isolates and STs that trace the majority of their ancestry to a single population, as well as admixed genotypes with significant ancestry from multiple populations. However, in concordance with the eBURST and ClonalFrame analyses, there is no large scale structuring by either site of isolation or geography. Most populations are present in roughly equal proportions in all disease and geographic region subgroups.

Population 2 (in red), comprised of the commensal *fucK* negative isolates, is an exception, with all but two isolates and STs coming from the US (all are commensal as well, of course). This cluster may represent a divergent population of NTHi with a significantly reduced ability to cause otitis media. While the majority of this population was collected from a single geographic region, it is possible that broader sampling techniques such as those used to collect the US commensal isolates (35, 128) would reveal additional members of this population in other regions, though this is pure speculation. Population 6(in orange) is another cluster that appears to be differentially distributed. Isolates with a high percentage of their ancestry from population 6 are more common among the OM isolates, particularly so among the Finland OM subgroup. This

may represent a case of combined geographic and disease population structuring, and would be an interesting target for further study.

The major differences between the analyses of the two datasets lie in populations 7 and 8 (brown and pink, respectively). Both populations are composed of multiple isolates of the same MLST genotype, ST57 for population 7 and ST34 for population 8. The identification of these groups of identical genotypes as populations by *structure* may be an artifact that reflects a violation of a model assumption, but has apparently had little detrimental effect on the proper clustering of other isolates and proved helpful in distinguishing these STs as being of interest, particularly ST57. The sole sequence type in population 7, ST57 was identified 18 times and was by far the most commonly identified genotype in this sample, occurring over three times more frequently than the next most common ST (which was ST34, containing five isolates). Additionally, every isolate of both ST57 and ST34 was recovered from the middle ear of a child with otitis media, and it was found in roughly equal proportions among the three geographic regions. Given that each of the ST57 isolates was also collected from a different child, this suggests a strong association between that particular genotype and otitis media, but very little stratification by geography. A similar trend for ST34 is seen, but the smaller number of isolates (five) precludes much hypothesizing. When analyzed as part of the unique STs dataset, both ST57 and ST34 were significantly admixed, tracing approximately 60% of their ancestry to population 4 and 30% to population 5. As it was not found at all among the commensal isolates, despite the naso- and oropharynges being the reservoir for

NTHi, ST57 may represent a rare genotype with increased virulence (thus, the OM collections would be ‘enriched’ for the virulent genotype). This genotype, along with ST34, has the potential to be a useful target for research into the mechanisms of NTHi virulence.

Chapter 4

Otitis Media Associated Polymorphisms in *Haemophilus influenzae*

INTRODUCTION

Association studies are among the most powerful techniques available to epidemiologists and health care researchers in discovering the determinants of disease. They present the closest approximation to the unobtainable ideal epidemiologic study, in which the frequency or probability of an outcome is assessed in the same population, both with and without the exposure of interest, simultaneously. Clearly this is an impossible situation, since a single population cannot be both exposed and unexposed at exactly the same time. As at least one of the exposure conditions is contrary to actual fact, this ideal, impossible study is known as a counterfactual contrast, and studies of association are designed to approximate it as closely as possible (117).

Studies of association have been popular and effective in identifying bacterial genetic determinants that have a statistically significant connection to a disease, thus implicating those loci in the virulence process. *H. influenzae* is among the many human pathogens studied with this technique, and a number of genes have been found to be associated with various diseases, including otitis media. Ecevit and colleagues examined the pilus gene cluster (*hifBC*) and several adhesins (*hmw1A, hmw2A, hmwC,* and *hia*) in Hib and NTHi by dot blot

hybridization and discovered a wide variability in prevalence of the genes (25). While *hifBC* (located in the hemagglutinating pili gene cluster) was significantly more prevalent among Hib isolates than among NTHi isolates, it was also more prevalent among commensal NTHi isolates than among OM NTHi isolates, presenting a complicated association with virulence. Furthermore, the *hmw* adhesin genes were more prevalent among OM NTHi isolates than among commensal NTHi isolates, suggesting they may have a role in the increased virulence of isolates collected from the middle ears of children with acute otitis media. A similar study by Pettigrew et al. screened 48 OM NTHi and 90 commensal NTHi isolates by dot blot hybridization and found that *lic2B* was found 3.7 times more frequently among the OM isolates (103). *lic2B* is a galactosyl transferase involved in biosynthesis of lipooligosaccharide (LOS), a strongly immunogenic, major component of the outer membrane. Presence of *lic2B* may lead to improved immune evasion by altering LOS structure. A 2007 publication by Juliao et al. identified the operon responsible for histidine biosynthesis (*his*) to be significantly more prevalent among OM NTHi isolates than among commensal NTHi isolates (55). By PCR and dot blot hybridization, the *his* operon was 62% more prevalent among the OM isolates, showing that metabolic pathways can also be virulence factors. Furthermore, the results also suggest that the environment in the middle ear may quite different than that of the throat and that the optimal fitness characteristics for survival and virulence of NTHi are different in those different environments.

Genes involved in the acquisition of iron and iron containing molecules

have also been the focus of association studies, which have implicated them in Hi virulence. Given the importance of iron for nearly all bacteria, and the absolute requirement of heme for *H. influenzae* aerobic growth, it is not surprising that Hi have several partially redundant systems to acquire iron from a variety of sources, including heme, hemoglobin, transferrin, hemoglobin:haptoglobin complexes, and heme:hemopexin complexes. A study by Morton et al. demonstrated that an mutant NTHi strain lacking the hemoglobin binding proteins (*hgp*'s) had significantly reduced virulence in an animal model of otitis media compared to the isogenic wild type strain (79). In a later study, Morton and colleagues showed that infant rats infected with a mutant Hib strain lacking the *hxuCBA* operon (responsible for the utilization of heme:hemopexin complexes) had significantly lower bacteremic titers and improved survival rates as compared to those infected with the wild type strain (82).

The hemin receptor of NTHi, *hemR*, has been the focus of considerably less research. *Haemophilus ducreyi*, a distant relative of *H. influenzae*, has a *hemR* homologue encoded by *tdhA*. Thomas et al. found that TdhA has significant sequence homology to HemR, as well as other heme receptors from gram negative bacteria, including HxuC from Hi, HmuR from *Yersinia pestis*, and ChuA from *E. coli* (131). Furthermore, an *E. coli* mutant unable to synthesize heme and lacking native heme and hemoglobin receptors but expressing *H. ducreyi tdhA* grew on low levels of heme only when an intact *H. ducreyi* Ton system plasmid was present, demonstrating functional TonB dependence. Leduc et al. found no statistically significant difference in pustule formation or

number of bacteria recovered from the pustules in six human volunteers experimentally inoculated with both wild type *H. ducreyi* and an isogenic *tdhA* mutant (66). These data led the authors to suggest that TdhA is not necessary for virulence in *H. ducreyi*.

The presence, however, of *hemR* in NTHi has been associated with otitis media. Xie and colleagues, using dot blot hybridization, found *hemR* to be significantly more common among OM NTHi isolates as compared to commensal NTHi isolates, yielding a prevalence ratio of 1.14 (146). Similarly, *hemR* was more common among invasive Hib isolates than among commensal NTHi isolates, with a prevalence ratio of 1.15. The hemin receptor was found in all Hib isolates and nearly all OM NTHi isolates, with only one of 121 OM NTHi isolates not hybridizing to the *hemR* specific probe. Recently, Whitby et al. used microarray and quantitative real-time PCR analyses to demonstrate that expression of *hemR* was increased under iron/heme limiting conditions in OM NTHi strain R2866, capsule deficient type d strain Rd, and invasive Hib strain 10810 (144).

Most genetic association studies in bacteria, including those described in the preceding paragraphs, do not attempt to control for the potentially confounding effect of population structure, despite numerous studies demonstrating the presence of structure in bacterial populations (21, 27, 32, 34, 91, 93, 94, 124, 126). The research presented in this chapter identifies amino acid polymorphisms in the NTHi hemin receptor HemR associated with otitis media while adjusting for the population structure identified in Chapter 3. In

addition, the theoretical three dimensional protein structure of HemR is assessed, allowing otitis media-associated HemR polymorphisms to be mapped onto potential functional and structural domains.

MATERIALS AND METHODS

DNA Amplification and Sequencing

The 170 confirmed NTHi isolates from Finland, Israel, and the US described in chapter three were utilized in this study. The PCR and DNA sequencing protocols used to analyze *hemR* were essentially identical to those used in Chapters 2 and 3 for MLST, except that the annealing temperatures were altered dependent on the primers used. All primers were designed using the fully sequenced NTHi strain 86-028NP as a template (47). As some primer pairs did not successfully amplify some isolates, alternate primers covering the same region of the gene were designed, yielding a total of seven primer pairs covering the entire coding region of *hemR* (**Table 4.1**). Sequences for the full coding region of *hemR* for all isolates used in this study are available from GenBank under accession numbers JN229266 through JN229411.

Data and Statistical Analysis

SeqMan Pro 8.1.5 (DNASTAR, Madison, WI) was used to align and trim the sequenced PCR products into the *hemR* open reading frame (ORF). Potential signal peptides were identified using the hidden Markov model implemented by the SignalP online server (4, 96). The predicted three

Table 4.1

	F Primer	R Primer	Ta[a]	Position[b]
Pair 1	AATTCGATGACTTGTTGTTTG	CCAGCATTTCTTACAGAACC	55	-153 to 633
Pair 2	TGCAAGAAAACCAGAAAATAG	ATATGGTCATAACGCACTCC	56	488 to 1316
Pair 2a	AATATGGCTGGCGGATTTC	TCACCCGAACTTGGTTTAGG	59	238 to 822
Pair 2b	TTCACCCGAGCATTTACTCAC	AGGTGCACGCCAAGTTCTAC	60	717 to 1486
Pair 3	TGGACGTCAATATACACAGG	GACCAATATTTTGCTTCAGG	56	1227 to 2052
Pair 3a	ACTATTATTTGGTTTGCAATGGT	CTTACAAACTCAGCACGCC	52	1110 to 2000
Pair 4a	AACGTGACACCTCACCAAGA	GTTATGATAAAGCCCGATTAGATTCA	56	1859 to +41

Primer pairs used to amplify *hemR*. NTHI strain 86-028NP was used as the template.
[a] Annealing temperature in degrees Celsius.
[b] Position of the resulting amplicons relative to *hemR* from strain 86-028NP. Numbers preceded by a - sign indicate positions upstream of the start codon, while numbers preceded by a + sign indicate positions downstream of the stop codon.

dimensional structures of the HemR protein from selected isolates were determined from their amino acid sequences by the I-TASSER server (118, 147, 148). Yasara View 11.6.16 was used to visualize the resulting theoretical protein structures (64).

All *hemR* sequences were translated and aligned in CLC Sequence Viewer 6.5.2 (CLC bio, Aarhus, Denmark); this amino acid sequence alignment was then examined visually to identify amino acid polymorphisms associated with otitis media. Unadjusted prevalence ratios (PRs), odds ratios (ORs), and associated confidence intervals (CIs) were calculated for the association between each potentially significant polymorphism and otitis media using OpenEpi 2.3.1 (20). All ratios compare OM isolates to commensal isolates. For all statistical analyses, isolates with the consensus amino acid at a polymorphic position based on the above alignment were coded with a '1', and isolates with any other amino acid at that position were coded with a '0'. Thus, a PR or OR

greater than one for a given polymorphism indicates that OM isolates are more likely to have the consensus amino acid at that position, while ratios less than one indicate a greater likelihood of having an alternative amino acid there.

To adjust the association between each HemR polymorphism and otitis media for the presence of population structure, a series of logistic regression models were constructed in R version 2.13.0 (111). Population structure information based on the all isolates dataset (as described extensively in Chapter 3) was used, and the estimated individual membership proportions (Q) at K=8 were coded as eight continuous variables (one for each population) with ranges between 0, indicating no ancestry from that population, and 100, indicating all ancestry was from that population. To obtain p-values for the ORs adjusted for population structure, permutation tests were constructed by randomly rearranging the polymorphism data labels and recalculating the OR 10,000 times. The proportion of permuted ORs as extreme or more extreme than the observed OR was the p-value for that permutation test.

RESULTS

HemR Characteristics

Sequencing of the full *hemR* protein coding sequence was attempted in the collection of 170 NTHi isolates characterized in Chapter 3 using the primers listed in **Table 4.1**. No amplification was observed with any of the *hemR* specific primers in ten isolates (one OM and nine commensal); these isolates were judged to be missing the gene entirely, which was not an unexpected result. A

previous study from this laboratory using dot blot hybridization to identify genes differentially distributed among OM and commensal NTHi found *hemR* to be more prevalent among OM isolates than among commensal isolates (99% versus 87%, PR = 1.14) (146). This study confirms those results, with a PR of 1.13 (95% CI = 1.03 – 1.25) comparing the prevalence of *hemR* in OM and commensal isolates. Intriguingly, all nine *hemR* negative commensal isolates were also *fucK* negative, suggesting that divergent NTHi, such as those that are *fucK* negative, are less likely to have potential virulence loci like *hemR*. However, the remaining five *fucK* negative commensal isolates, as well as the sole OM *fucK* negative isolate, were positive for *hemR* so the connection is by no means absolute. A further 14 isolates had consistently poor or absent amplification for at least one of the primer pairs that could not be resolved by using fresh reagents (including new purified genomic DNA) or altering the PCR annealing temperature. As the full ORF of the gene could not be determined, these isolates were excluded from further analyses. Thus, the full protein coding sequence of *hemR* was obtained for a total of 146 NTHi isolates. A summary of these results appears in **Table 4.2**.

The nucleotide sequences for all 146 isolates were translated and aligned using CLC Sequence Viewer 6.5.2. The length of the HemR protein was quite variable between isolates, with an average of 744.5 residues, a high of 751 residues, and a low of 724 residues. The isolate with the shortest HemR at 724 amino acids (K35LE2.1, a US OM isolate) was an outlier, however, as the next shortest HemR contained 740 residues. This outlier had little effect on

Table 4.2

	Initial[a]	Gene Missing[b]	Poor Amplification[c]	Total Removed	Final
Finland OM	30	0	3	3	27
Finland Commensal	26	0	3	3	23
Israel OM	28	0	5	5	23
Israel Commensal	22	1	2	3	19
US OM	37	1	1	2	35
US Commensal	27	8	0	8	19
Total OM	95	1	9	10	85
Total Commensal	75	9	5	14	61

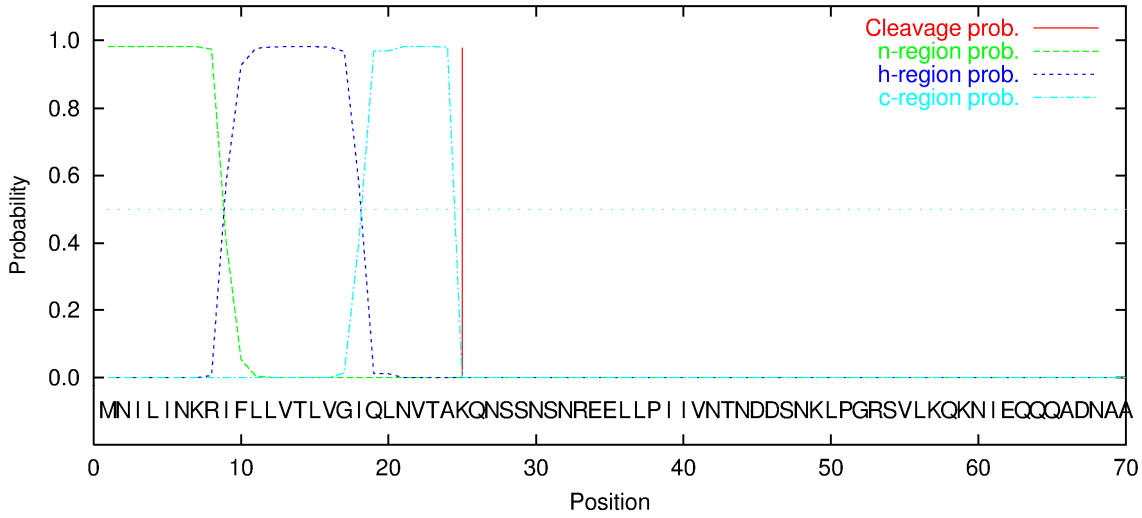
Summary of the NTHi isolates used for *hemR* amplification.
[a] Initial number of NTHi isolates characterized in Chapter 3.
[b] Isolates in which *hemR* is not present.
[c] Isolates in which full sequencing of *hemR* was not possible due to poor PCR amplification.

calculating the average, as both the median and the mode number of amino acids were 745, essentially the same as the average. The majority of the insertions and deletions responsible for the varying length occurred in two locations, approximately at positions 510 and 580 relative to the N-terminus.

Most bacterial outer membrane proteins contain a short peptide chain, often near the N-terminus, that assists in directing the transport of the protein through the periplasm. As HemR is an outer membrane protein, the presence of a signal peptide was assessed using the hidden Markov model implemented in the SignalP prediction server (4, 96). Based on the HemR alignment, four NTHi isolates (F162-7, F433-3, I207, and C09.2.3) that adequately covered the small amount of variation observed near the N-terminus were chosen for analysis. An extremely high probability for the presence a signal peptide was observed in all four isolates (mean ± SD = 96.5% ± 1.8%). The analyses agreed on the

cleavage site being between residues 24 and 25, as well (96.1% ± 1.9%). Similarly high probabilities were estimated for the presence of a canonical n-region (a positively charged stretch of amino acids), h-region (the hydrophobic core area), and c-region (a polar amino acid stretch that includes the cleavage site). A representative plot of the SignalP results for isolate F433-3 is shown in **Figure 4.1**. It is not known, however, whether this putative signal peptide is indeed cleaved from the mature protein.

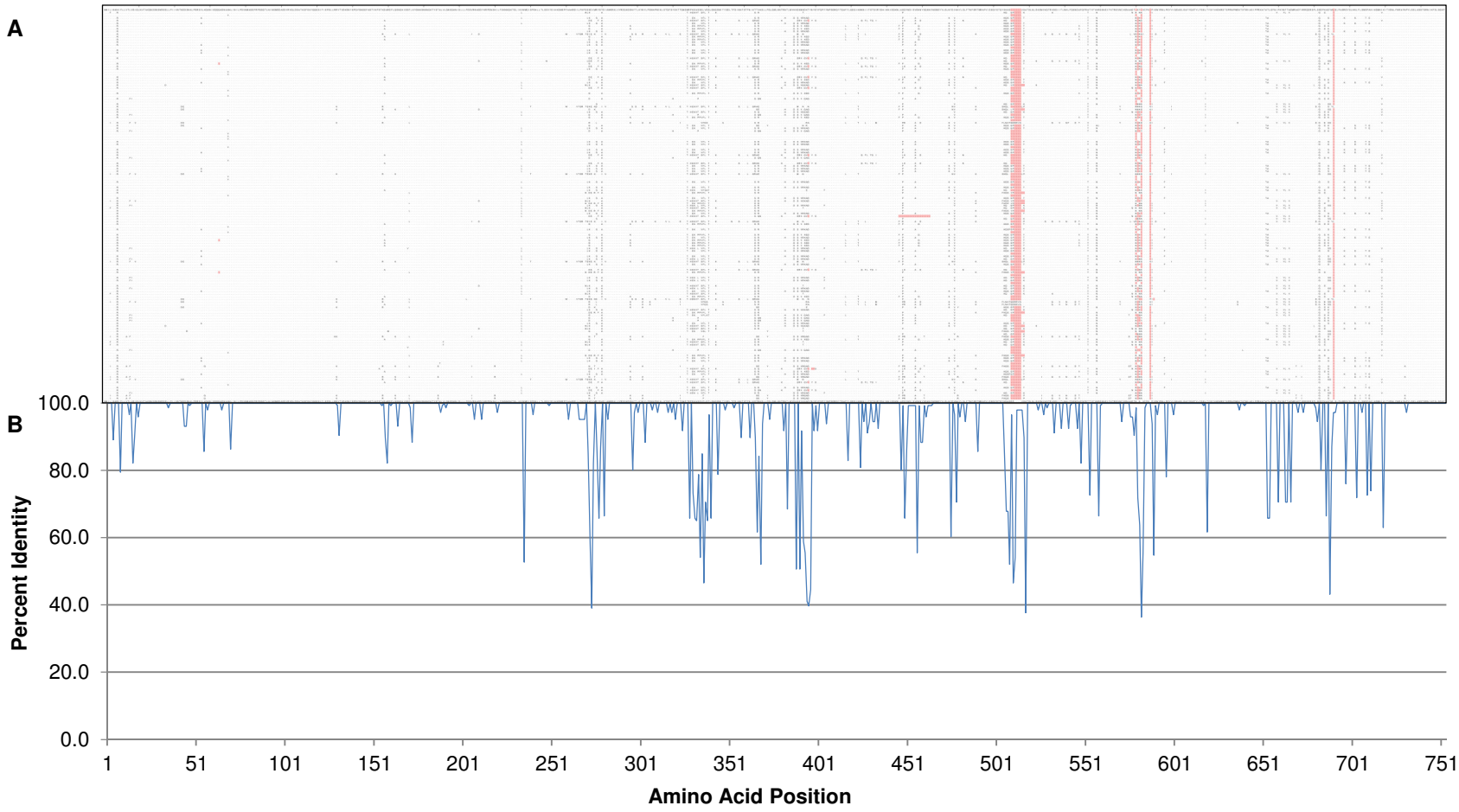
Figure 4.1



Signal peptide probabilities estimated by SignalP. HemR from isolate F433-3 was used, and position zero is the N-terminus. The n-, h-, and c-regions are canonical in most signal peptides and are a stretch of positively charged amino acids, a hydrophobic area, and a polar region, respectively.

The overall level of amino acid sequence conservation was quite high, with an average of 94.5% identity. However, the level of conservation fluctuated considerably across the length of the protein, with some regions identical in all 146 isolates while others exhibited substantial variation. This can clearly be seen in the plot of percent identity versus amino acid position shown in **Figure 4.2**. The N-terminal 250 amino acids, approximately 1/3 of the protein,

Figure 4.2



HemR amino acid sequence conservation by position. The total length of the alignment, including gaps, is 753 residues. **A**. Alignment of the 146 NTHi HemR amino acid sequences. Identical residues are coded with a dot, and gaps are colored pink. **B**. Plot of the percent identity by amino acid position.

shows considerable conservation, with only one position at lower than 80% identity. The picture is remarkably different after residue 250, where regions of substantial diversity (40 – 75% identity) are interspersed with regions of high conservation. Unsurprisingly, this pattern is directly related to the different functional domains and motifs of the protein, as will be further explored in subsequent sections.

In all but four cases, multiple isolates with the same MLST genotype also shared identical HemR sequences. Isolates of STs 13 and 156 had single non-synonymous SNPs resulting in alterations at positions 171 and 235, respectively. Of the two isolates comprising ST11, one was missing the entire gene, as mentioned previously. Finally, two of the three isolates of ST3 had identical HemR sequences, while the third showed only 94.5% sequence identity to the other two. As the residue differences were spread throughout the sequence and not concentrated within one PCR amplicon, it seems likely that this isolate's HemR sequence is legitimately different from other members of its MLST genotype, potentially due to a fairly recent recombination event.

Based on amino acid identity, there were 47 unique HemR protein sequences (differed by at least one amino acid, and henceforth referred to as HemR 'types') among the 146 NTHi isolates. While the majority of HemR types consisted of only a single isolate, six types were comprised of at least five different STs. The largest, HemR type 4 (arbitrarily named), contains 19 isolates of 11 different STs and consists of a mix of OM and commensal isolates from all three geographic regions, a theme that is repeated amongst most of the other

multi-ST HemR types. This pattern is reminiscent of the population structure inferred in Chapter 3, and adding the population structure information from the all isolates dataset to the HemR types clarifies the picture somewhat. **Table 4.3** sorts the HemR types, and the isolates they contain, by the individual membership proportion (*Q*) for each isolate. It is apparent that for the most part, when HemR types are comprised of isolates of multiple STs, those isolates belong to the same population (for example, HemR type 4 in population 2 and HemR type 2 in population 6). This is not surprising, as isolates within a population are likely to be more related and/or exchange their genetic material more frequently with each other than with isolates in other populations.

Otitis Media Associated HemR Polymorphisms

Visual assessment of the HemR alignment identified 14 amino acid polymorphisms significantly associated with otitis media at the 95% confidence level, with a further four polymorphisms significant at the 90% confidence level (**Table 4.4**). The name of each polymorphism references its position within the HemR alignment; positions within specific sequences will vary somewhat. Four of these polymorphisms (I659V, N663Y, F664L, and A666V) always co-occurred and are treated as a single entity for the statistical analyses, leaving a total of 15 polymorphisms significant at the 90% confidence level. While prevalence ratios are perhaps the most intuitive and straightforward measure of the crude association between otitis media and these polymorphisms, the equivalent ORs are also reported in **Table 4.4** for comparison purposes, as later analyses rely on

Table 4.3

Population 1					Population 2							Population 3						
Isolate	ST	1° Pop[a]	% [b]		Isolate	ST	1° Pop[a]	% [b]	2° Pop[c]	% [d]		Isolate	ST	1° Pop[a]	% [b]	2° Pop[c]	% [d]	
HemR Type 30					**HemR Type 4**							**HemR Type 8**						
C09.2.3	10006	1	99.5%		F571	159	2	99.5%				I191	238	3	98.9%			
P11.2.3	10006	1	99.5%		F1897-7	159	2	99.5%				I218	238	3	98.9%			
P20.2.1	10010	1	99.3%		F894	155	2	99.4%				G723	885	3	98.4%			
					54.3-22	155	2	99.4%				I278	863	3	84.5%			
HemR Type 38					F206-4	155	2	99.4%				I175	871	3	78.3%	7	20.0%	
I328	10001	1	98.2%		F286-5	155	2	99.4%				I280	892	5	49.6%	3	48.4%	
					G1322	633	2	99.4%										
HemR Type 32					M02.2.4	414	2	82.1%	5	16.4%		**HemR Type 35**						
F05.2.3	10008	1	92.3%		F1831	395	2	74.0%	6	25.1%		28.4-21	176	3	99.1%			
					63.4-22	196	2	72.9%	6	22.2%								
					I283	165	2	72.8%	5	25.4%		**HemR Type 37**						
					I153	165	2	72.8%	5	25.3%		I338	867	3	84.4%			
					F441	253	2	68.2%	6	30.9%		I340	867	3	84.4%			
					I213	244	2	50.0%	3	20.6%								
					F1268-9	36	5	79.0%	2	19.5%		**HemR Type 43**						
					F1388-5	36	5	78.9%	2	19.5%		I255	890	3	84.3%			
					F1449-9	36	5	78.9%	2	19.5%								
					G522	36	5	78.8%	2	19.6%		**HemR Type 17**						
					G1222	887	5	67.4%	2	30.5%		K16RE2.4	584	3	75.3%			
					HemR Type 9							**HemR Type 42**						
					P20LE2.10	107	2	98.1%				I270	891	3	62.6%	6	18.9%	
					E12.2.11	107	2	98.0%										
					HemR Type 19							**HemR Type 34**						
					I210	878	2	78.0%				65.2-23	170	3	59.9%	7	31.7%	
					HemR Type 40							**HemR Type 18**						
					F1448	264	2	48.0%	7	28.9%		F162-7	411	3	59.7%	7	39.0%	
					I307	264	2	47.9%	7	29.0%		F443	411	3	59.6%	7	39.0%	
												F994	411	3	59.6%	7	39.0%	
												F1115	411	3	59.6%	7	39.1%	
					HemR Type 47							I179	411	3	59.6%	7	39.1%	
					F1983	889	2	36.8%	3	32.3%		F1308	855	3	56.8%	7	34.2%	
												K7LE2.5	10011	3	55.8%	2	24.1%	
												HemR Type 23						
												I183	872	3	53.1%	2	41.4%	
												HemR Type 22						
												I184	873	3	46.3%	2	44.8%	

- Commensal
- OM
- Finland
- Israel
- US

HemR types sorted by the largest membership proportion (*Q*) for each isolate. The HemR types are arbitrarily numbered 1 - 47. Isolates are color coded by disease and geographic region. Brown shaded cells indicate isolates whose primary population differs from the majority within that HemR type.

[a] Primary population from whence an isolate acquired the largest fraction of its ancestry.
[b] Membership proportion in the primary population.
[c] Secondary population in which an isolate has the second largest membership (must be ≥15%).
[d] Membership proportion in the secondary population.

Table 4.3 continued

Population 4						Population 5					
Isolate	ST	1° Pop[a]	%[b]	2° Pop[c]	%[d]	Isolate	ST	1° Pop[a]	%[b]	2° Pop[c]	%[d]
HemR Type 45						**HemR Type 3**					
I247	311	4	90.6%			F324	13	5	99.0%		
HemR Type 6						**HemR Type 11**					
G06.2.1	594	4	85.2%			I316	204	5	99.0%		
G1423	594	4	85.1%			I181	204	5	99.0%		
F651	852	4	52.6%	3	22.5%	P5ME2.5	204	5	99.0%		
HemR Type 20						**HemR Type 15**					
I207	895	4	84.8%	7	14.4%	F1248	13	5	99.0%		
HemR Type 16						O07.2.12	149	5	99.0%		
K26LE2.7	683	4	79.9%			K29RE2.10	112	5	99.0%		
HemR Type 13						F1158	163	5	98.9%		
K17ME2.5	146	4	68.9%	2	28.7%	F638	163	5	98.8%		
37.3-21	146	4	68.8%	2	28.8%	F1450	888	5	78.9%		
K27RE2.8	146	4	68.8%	2	28.8%	**HemR Type 25**					
K33RE2.4	146	4	68.8%	2	28.7%	I289	393	5	87.0%		
F1060	260	4	63.8%	6	28.4%	I156	393	5	87.0%		
HemR Type 44						**HemR Type 21**					
I249	859	4	63.9%	2	21.3%	I264	861	5	67.0%	3	24.9%
						I188	876	5	52.8%	6	35.9%
						HemR Type 39					
						I312	865	5	64.9%	6	16.4%
						HemR Type 27					
						F885-7	838	5	64.1%	6	34.8%
						HemR Type 24					
						I162	868	5	60.2%	6	19.1%
						HemR Type 12					
						K35LE2.1	883	5	60.0%	3	35.3%
- Commensal						N02.2.3	881	3	45.3%	2	40.8%
- OM						**HemR Type 46**					
						I242	166	5	57.0%	6	35.6%
- Finland						**HemR Type 10**					
- Israel						P6RE2.9	182	5	53.8%	6	43.7%
- US						**HemR Type 41**					
						I276	862	5	52.5%	6	36.5%

[a] Primary population from whence an isolate acquired the largest fraction of its ancestry.
[b] Membership proportion in the primary population.
[c] Secondary population in which an isolate has the second largest membership (must be ≥15%).
[d] Membership proportion in the secondary population.

Table 4.3 continued

Population 6

Isolate	ST	1° Pop[a]	%[b]	2° Pop[c]	%[d]
HemR Type 29					
F1942	3	6	99.4%		
F164-6	3	6	99.4%		
HemR Type 2					
P26RE2.2	3	6	99.4%		
I336	866	6	99.3%		
F242	33	6	99.1%		
D07.2.6	33	6	99.1%		
F2206-6	137	6	96.5%		
F433-3	40	6	96.4%		
F1152-8	40	6	96.4%		
F608-5	12	6	72.7%	8	26.4%
F1124-2	12	6	72.7%	8	26.4%
I168	12	6	72.7%	8	26.4%
P16LE2.7	12	6	72.7%	8	26.4%
F202	160	6	71.8%	2	15.0%
F1702-5	160	6	71.8%	2	15.3%
F282	851	6	68.6%	8	25.7%
I198	877	6	49.1%	3	48.7%
HemR Type 14					
K32RE2.5	143	6	99.3%		
HemR Type 7					
F1618-4	259	6	94.3%		
G922	203	6	92.2%		
F1726-1	147	6	79.1%		
F1015	245	6	76.4%	2	22.8%
F251-1	245	6	76.4%	2	22.9%
F2188-8	243	6	55.3%	5	19.3%
H04.2.1	11	6	45.6%	4	38.1%
G622	884	6	41.2%	5	23.0%
HemR Type 33					
C17.2.11	879	6	50.1%	5	45.5%
HemR Type 31					
J06.2.2	880	6	49.0%	5	25.9%
30.2-24	2	6	39.1%	7	32.6%
HemR Type 26					
F1663-8	43	6	37.7%	3	29.2%
No HemR					
K15RE2.6	11	6	45.6%	4	38.2%

Population 7

Isolate	ST	1° Pop[a]	%[b]
HemR Type 5			
F199-3	57	7	99.6%
F1296-5	57	7	99.6%
F1588-3	57	7	99.6%
F2025-4	57	7	99.6%
F2037-2	57	7	99.6%
I164	57	7	99.6%
I169	57	7	99.6%
I202	57	7	99.6%
I221	57	7	99.6%
I224	57	7	99.6%
I226	57	7	99.6%
K19RE2.4	57	7	99.6%
P19LE2.6	57	7	99.6%
P25RE2.7	57	7	99.6%
G322	57	7	99.6%
G822	57	7	99.6%
G1522	57	7	99.6%
G1623	57	7	99.6%
HemR Type 36			
I345	84	7	72.6%

Population 8

Isolate	ST	1° Pop[a]	%[b]	2° Pop[c]	%[d]
HemR Type 1					
F1232-4	34	8	99.5%		
G123	34	8	99.5%		
G423	34	8	99.5%		
F1123	34	8	99.5%		
F1822	34	8	99.5%		
K8RE2.1	156	8	48.8%	6	35.3%
K21LE2.7	882	3	52.5%	8	36.4%
K34LE2.1	882	3	52.5%	8	36.3%
I230	396	3	51.4%	8	36.8%
F120	472	3	44.1%	8	43.3%
HemR Type 28					
F658-2	156	8	48.8%	6	35.6%

- Commensal
- OM
- Finland
- Israel
- US

[a] Primary population from whence an isolate acquired the largest fraction of its ancestry.
[b] Membership proportion in the primary population.
[c] Secondary population in which an isolate has the second largest membership (must be ≥15%).
[d] Membership proportion in the secondary population.

logistic regression models that output ORs, and there is unfortunately no straightforward comparison between the two in most situations. Consequently, it is important to note the difference in interpretation between PRs and ORs, given the large differences in value for the same relationship. As an example, consider the first polymorphism in **Table 4.4**, L4I/F. The PR is 1.15, indicating that the prevalence of the consensus amino acid leucine at position four of HemR is 1.15 times (or 15%) greater among OM isolates than among commensal isolates. The OR for that same relationship is much greater (3.52) because it is measuring something different. In this case, an OR of 3.52 indicates that the odds of an OM isolate having a leucine at position four are 3.52 times higher than the odds for a commensal isolate. When the outcome (a unique polymorphism, in this study) is rare, the OR will usually approximate the PR, but this is clearly not the case here. Both the regular p-value and a p-value calculated by a permutation test are reported for the ORs, again for comparison purposes. The permutation tests were performed by randomly rearranging the polymorphism data labels and recalculating the OR 10,000 times. The p-values were calculated as the proportion of permuted ORs as extreme or more extreme than the observed OR.

All of the significant PRs presented in **Table 4.4** are modest in magnitude, with largest at 1.41 for the P589I polymorphism. This is not unexpected, as both truly commensal NTHi and NTHi capable of causing disease are found in the naso- and oropharynges, which can have a diluting effect on ratio measures of association. Of the 15 identified polymorphisms, 13 have PRs greater than the null value of one, signifying that OM isolates are significantly more likely to have

Table 4.4

Polymorphism			OM		Commensal						
1° aa[a]	Position	2° aa[b]	$N_{1°}/N_{total}$[c]	%	$N_{1°}/N_{total}$[c]	%	PR (95% CI)	p-value	OR (95% CI)	p-value	Perm. p-value[d]
L	4	I/F 11/5	80/85	94.1%	50/61	82.0%	1.15 (1.01 - 1.31)	0.02	3.52 (1.15 - 10.73)	0.03	0.02
R	8	H	63/85	74.1%	53/61	86.9%	0.85 (0.73 - 1.00)	0.06	0.43 (0.18 - 1.05)	0.06	0.04
L	15	F	77/85	90.6%	43/61	70.5%	1.29 (1.08 - 1.53)	<0.01	4.03 (1.62 - 10.04)	0.00	<0.01
V	16	I	80/85	94.1%	51/61	83.6%	1.13 (1.00 - 1.27)	0.04	3.14 (1.01 - 9.71)	0.05	0.03
A	70	V	67/85	78.8%	59/61	96.7%	0.82 (0.72 - 0.92)	<0.01	0.13 (0.03 - 0.57)	0.01	<0.01
R	131	K	81/85	95.3%	51/61	83.6%	1.14 (1.01 - 1.29)	0.02	3.97 (1.18 - 13.33)	0.03	0.02
K	157	R	80/85	94.1%	51/61	83.6%	1.13 (1.00 - 1.27)	0.04	3.14 (1.01 - 9.71)	0.05	0.03
Q	164	K	82/85	96.5%	54/61	88.5%	1.09 (0.99 - 1.20)	0.06	3.54 (0.88 - 14.30)	0.08	0.06
N	303	H/R 13/4	79/85	92.9%	50/61	82.0%	1.13 (0.99 - 1.29)	0.04	2.90 (1.01 - 8.33)	0.05	0.03
T	324	I/E 8/4	81/85	95.3%	53/61	86.9%	1.10 (0.98 - 1.22)	0.07	3.06 (0.88 - 10.66)	0.08	0.07
Y	405	F	83/85	97.6%	54/61	88.5%	1.10 (1.00 - 1.21)	0.02	5.38 (1.08 - 26.87)	0.04	0.02
D	578	N	50/85	94.1%	52/61	85.2%	1.10 (0.98 - 1.24)	0.07	2.77 (0.88 - 8.73)	0.08	0.07
P	589	I	53/85	62.4%	27/61	44.3%	1.41 (1.02 - 1.95)	0.03	2.09 (1.07 - 4.07)	0.03	0.02
I*	659*	V*	67/85	78.8%	36/61	59.0%	1.34 (1.05 - 1.69)	0.01	2.58 (1.25 - 5.36)	0.01	0.01
N*	663*	Y*									
F*	664*	L*									
A*	666*	V*									
I	718	V	60/85	70.6%	32/61	52.5%	1.35 (1.02 - 1.77)	0.03	2.18 (1.10 - 4.32)	0.03	0.02

Distribution of HemR polymorphisms among OM and commensal NTHi isolates.

[a] Consensus amino acid at that position.

[b] Alternative amino acid at that position. Polymorphisms at positions 4, 303, and 324 have multiple alternative amino acids.

[c] Prevalence of the consensus (or 1°) amino acid at a given position within OM or commensal isolates.

[d] P-values calculated by a permutation test.

* The I659V, N663Y, F664L, and A666V polymorphisms always co-occur, and are thus treated as a single entity in the statistical analyses.

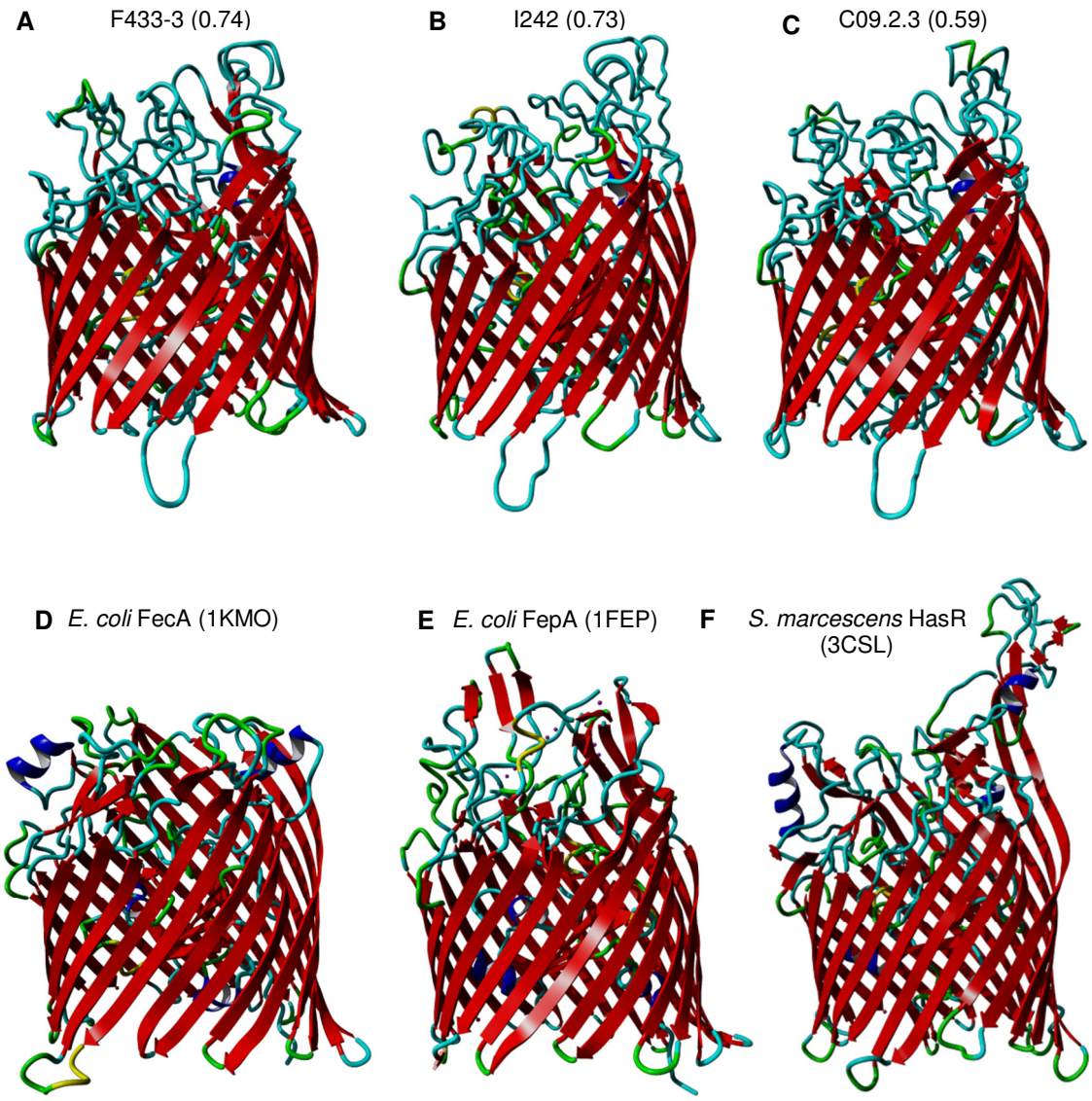
the consensus amino acid at those positions within HemR. This suggests a constraint on the diversity among these isolates, consistent with the idea that NTHi isolates able to cause otitis media differ in some ways from those that typically do not. It is possible that alterations at certain loci like HemR can modify the virulence potential of an isolate, and that as one would expect, most alterations either provide no benefit or are deleterious.

HemR Protein Structure

To assist in identifying the potential functional effects of these polymorphisms, the three dimensional tertiary structure of the full HemR protein was predicted using the I-TASSER protein structure and function server (118, 147, 148). I-TASSER is highly regarded, and was ranked as the top server for protein structure prediction in the last three Critical Assessment of Techniques for Protein Structure Prediction (CASP) competitions. The confidence score, or C-score, is a statistic calculated by I-TASSER for estimating the quality of the predicted models. It typically has a range of -5 to 2, such that a higher value signifies a model with high confidence. Yasara View 11.6.16 was used to visualize the resulting theoretical protein structures (64).

Figure 4.3 shows the predicted structure for HemR from three NTHi isolates in panels A – C, along with their respective C-scores. Each HemR sequence was of a different HemR type. All three models have high confidence based on their C-scores, which are close to the maximum value of 2. Panels D – F show the experimentally derived (via X-ray diffraction) structures of three iron

Figure 4.3



Predicted three dimensional structure of HemR. Secondary structures are color coded as follows: red - β-strands; dark blue – α-helices; green – turns; light blue – coils. All models are oriented with the extracellular side on top.

A-C. Theoretical structures for HemR predicted by I-TASSER. The C-score for each model is in parentheses. **A**. Finland OM isolate F433-3. **B**. Israel commensal isolate I242. **C**. US commensal *fucK* (-)ve isolate C09.2.3.

D-F. Experimentally derived structures for three iron acquisition receptors in other gram (-) species. The PDB IDs are in parentheses. **D**. *E. coli* ferric citrate uptake transporter FecA. **E**. *E. coli* ferric enterobactin receptor FepA. **F**. *S. marcescens* hemophore receptor HasR.

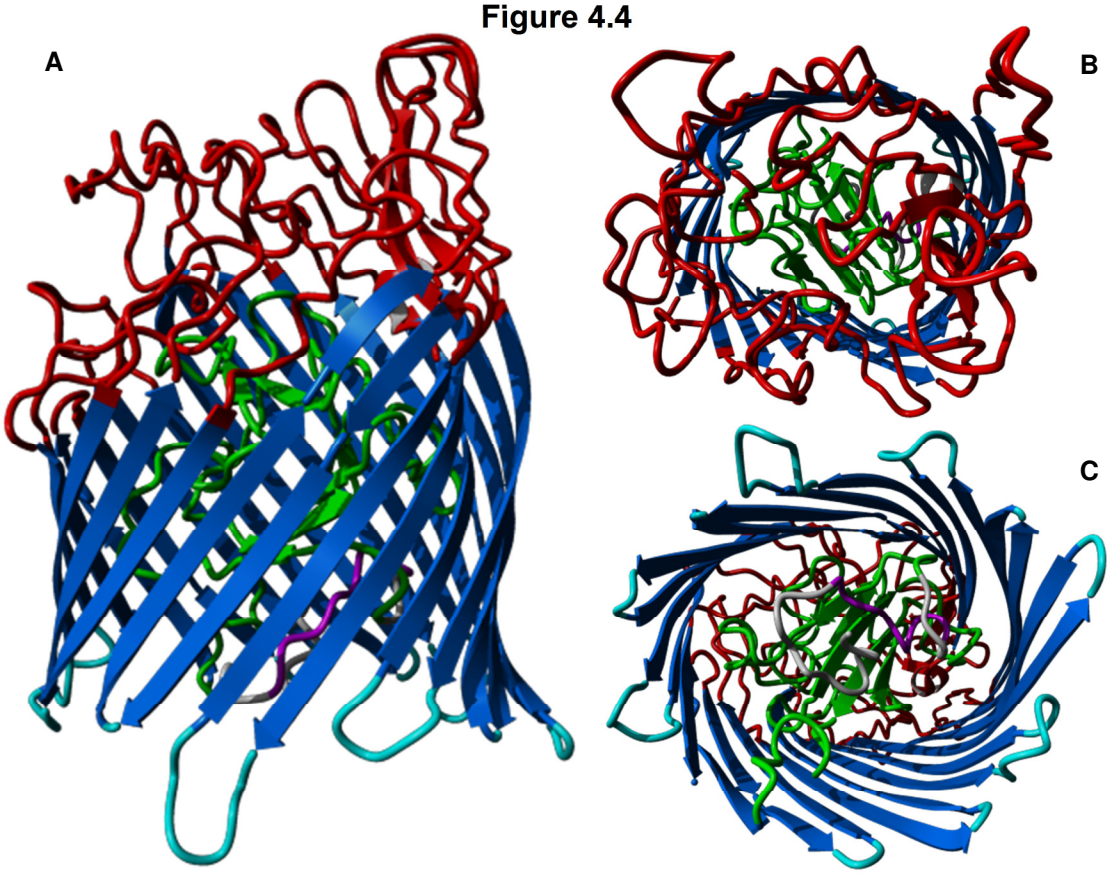
acquisition receptors from other gram negative species. There is marked similarity between the predicted structures for NTHi HemR and the known structures from other bacteria, despite very low amino acid sequence identity (approximately 18%, 15%, and 22% between F433-3 and FecA, FepA, and HasR, respectively). HemR appears to be a member of a common class of TonB dependent, ligand-gated channels formed by a monomeric, 22 strand, anti-parallel beta-barrel. Many highly specific, high affinity outer membrane receptors, including many iron acquisition proteins, fall into this category, including many *E. coli* receptors and the heme:hemopexin binding protein C (HxuC) and the transferrin binding protein 1 (Tbp1) in Hi (predicted structures; data not shown). Typically, the N-terminal 150 – 200 residues form a plug domain in the periplasmic end of the beta-barrel, blocking diffusion through the receptor. Binding of the ligand to the extracellular domains and TonB to the periplasmic side induces a conformational alteration of the channel, allowing passage of the ligand. Given the remarkable similarity between the structures of HemR and the known TonB dependent receptors, there is little doubt that HemR functions in the same manner.

TonB dependent receptors also have two relatively conserved domains, one at the N-terminus and the other at the C-terminus. The N-terminal domain, known as the TonB box, is involved in the interaction with TonB. PROSITE (http://prosite.expasy.org) lists the TonB box pattern as <x(10,115)-[DENF]-[ST]-[LIVMF]-[LIVSTEQ]-V-[AGPN]-[AGP]-[STANEQPK]; a key feature appears to be the invariant valine in the midst of the domain (PROSITE accession number

PS00430). There is a short sequence in HemR beginning 37 residues from the N-terminus (LPIIVNTN) that is a partial match (five of eight residues, including the invariant valine) with the TonB box pattern. This may be the equivalent domain for this protein in NTHi. The C-terminus domain (PROSITE accession number PS01156) is marked by two invariant residues; the HemR sequence in this region matches exactly to the listed pattern

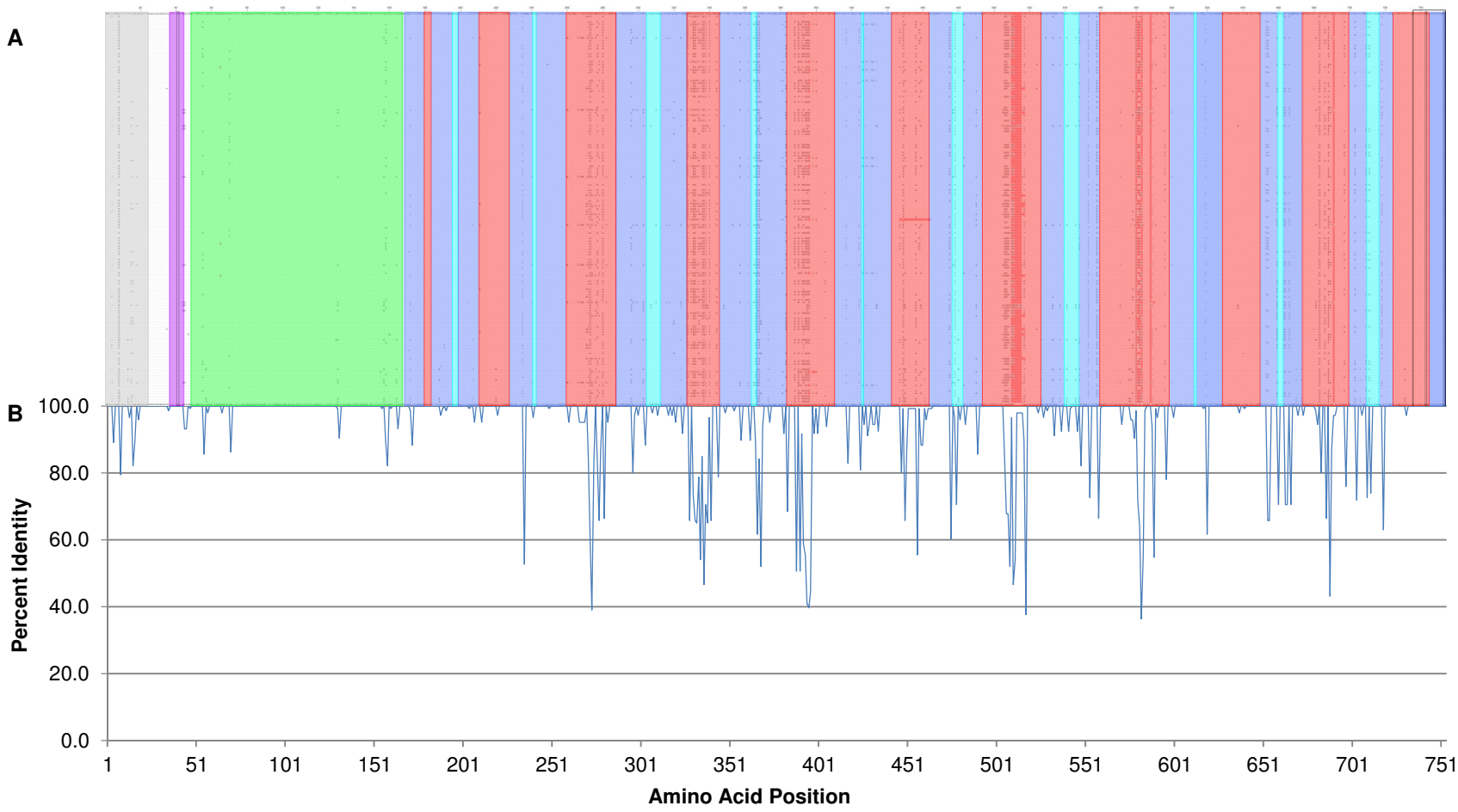
Knowing the theoretical three dimensional structure of HemR allows a more detailed assessment of the variable amino acid conservation observed in this sample. **Figure 4.4** shows a structural model of HemR with various domains

Figure 4.4



Predicted structure of HemR from isolate F433-3. Structural domains are colored as follows: gray – signal peptide; purple – putative TonB box; green – N-terminal plug domain; dark blue – transmembrane β-strands; red – extracellular loops; light blue – intracellular loops. **A**. Side view. **B**. Extracellular side. **C**. Periplasmic side.

Figure 4.5

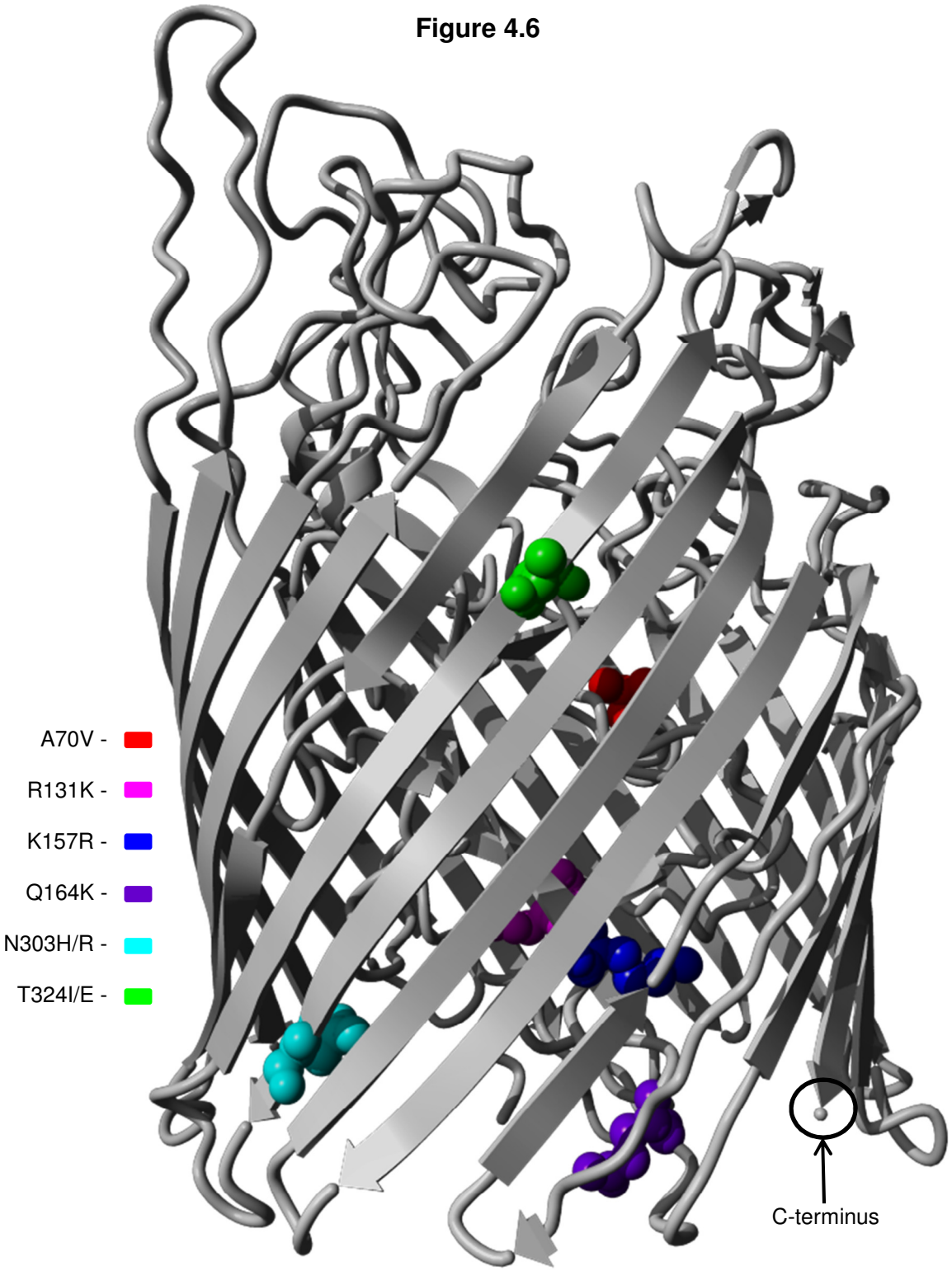


HemR amino acid sequence conservation by structural domain. **A**. Alignment of the 146 NTHi HemR amino acid sequences. Identical residues are coded with a dot. Structural domains match those in Figure 4.4 and are colored as follows: gray – signal peptide; purple – putative TonB box; green – N-terminal plug domain; dark blue – transmembrane β-strands; red – extracellular loops; light blue – intracellular loops. **B**. Plot of the percent identity by amino acid position.

color coded, including extra- and intracellular loops and the N-terminal plug domain. By mapping these structural domains onto the alignment of all 146 NTHi HemR sequences, it becomes readily apparent that much of the variation in sequence conservation corresponds to these different domains (**Figure 4.5**). The majority of the most variable regions map to extracellular loops, as one would expect for an outer membrane protein exposed to the immune system. Interestingly, four of the extracellular loops appear to be quite conserved; they may be shielded from the immune system and thus have little pressure to vary, or they may be involved in ligand binding and be functionally constrained. Likewise, most of the intracellular loops are fairly conserved, but several show relatively high levels of variation.

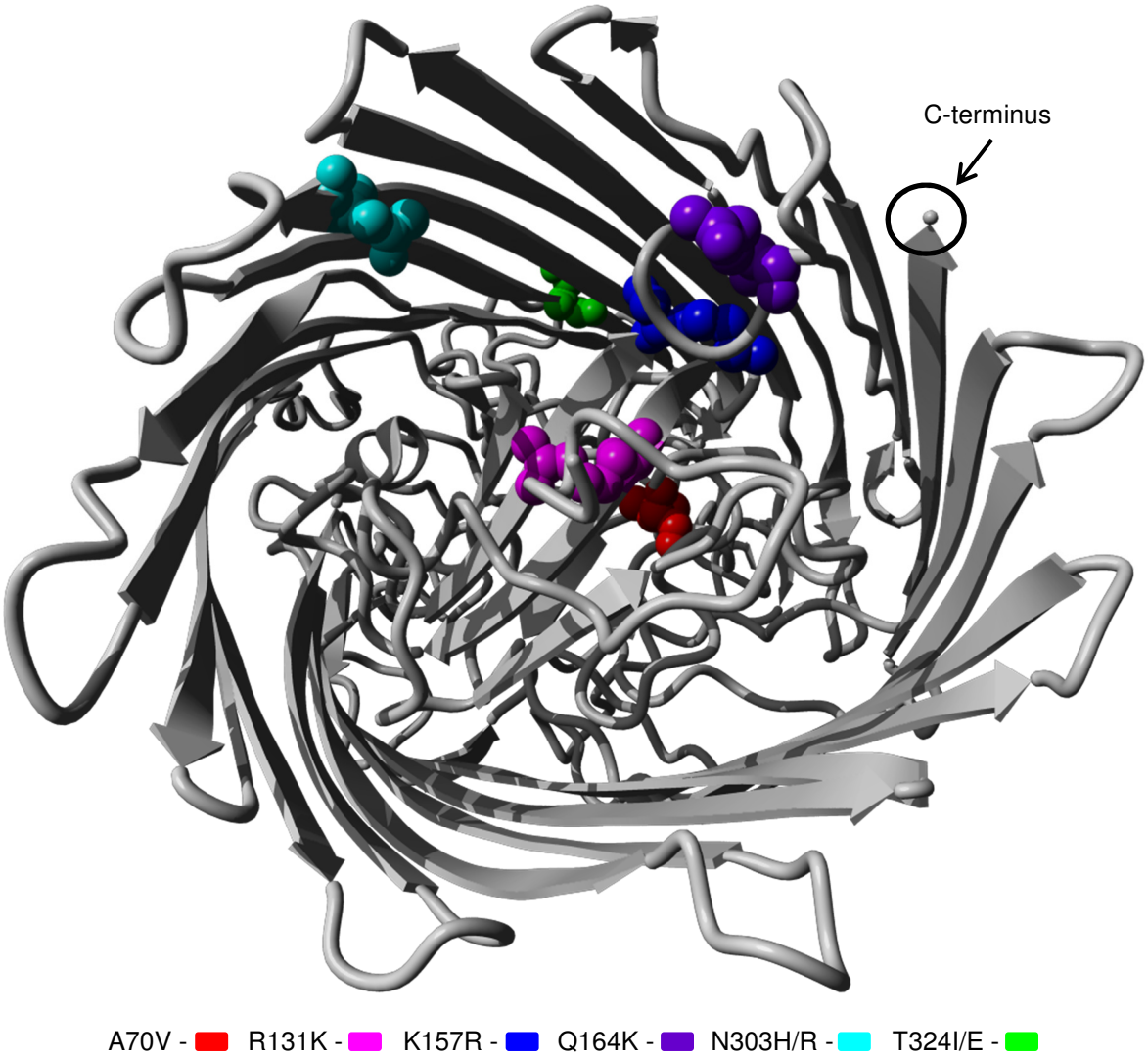
Prediction of the three dimensional structure of HemR also allows more informative hypotheses of the potential functional effect of the otitis media-associated polymorphisms. **Figures 4.6 & 4.7** highlight the A70V, R131K, K157R, Q164K, N303H/R, and T324I/E polymorphisms in two different views of HemR, while **Figure 4.8** depicts the Y405F, D578N, P598I, and I718V polymorphisms, as well as the I659V, N663Y, F664L, and A666V quartet of polymorphisms that are treated as a single entity. All three figures have the C-terminus circled in black to aid the viewer in orientation between them. Furthermore, all figures are modeled on the HemR structure from Finland OM isolate F199-3 of ST57 and population 7. The HemR sequences from the ST57 isolates are useful models in that they have the OM-associated amino acid at all the polymorphic positions identified in **Table 4.4** (i.e. if the PR > 1 they have the

Figure 4.6



Side view of the predicted structure of HemR from ST57 isolate F199-3 showing the locations of polymorphisms A70V through T324I/E. The N-terminal signal peptide has been removed in this figure. The six polymorphisms are color coded as shown and have been depicted using a space-filling model for added visibility.

Figure 4.7

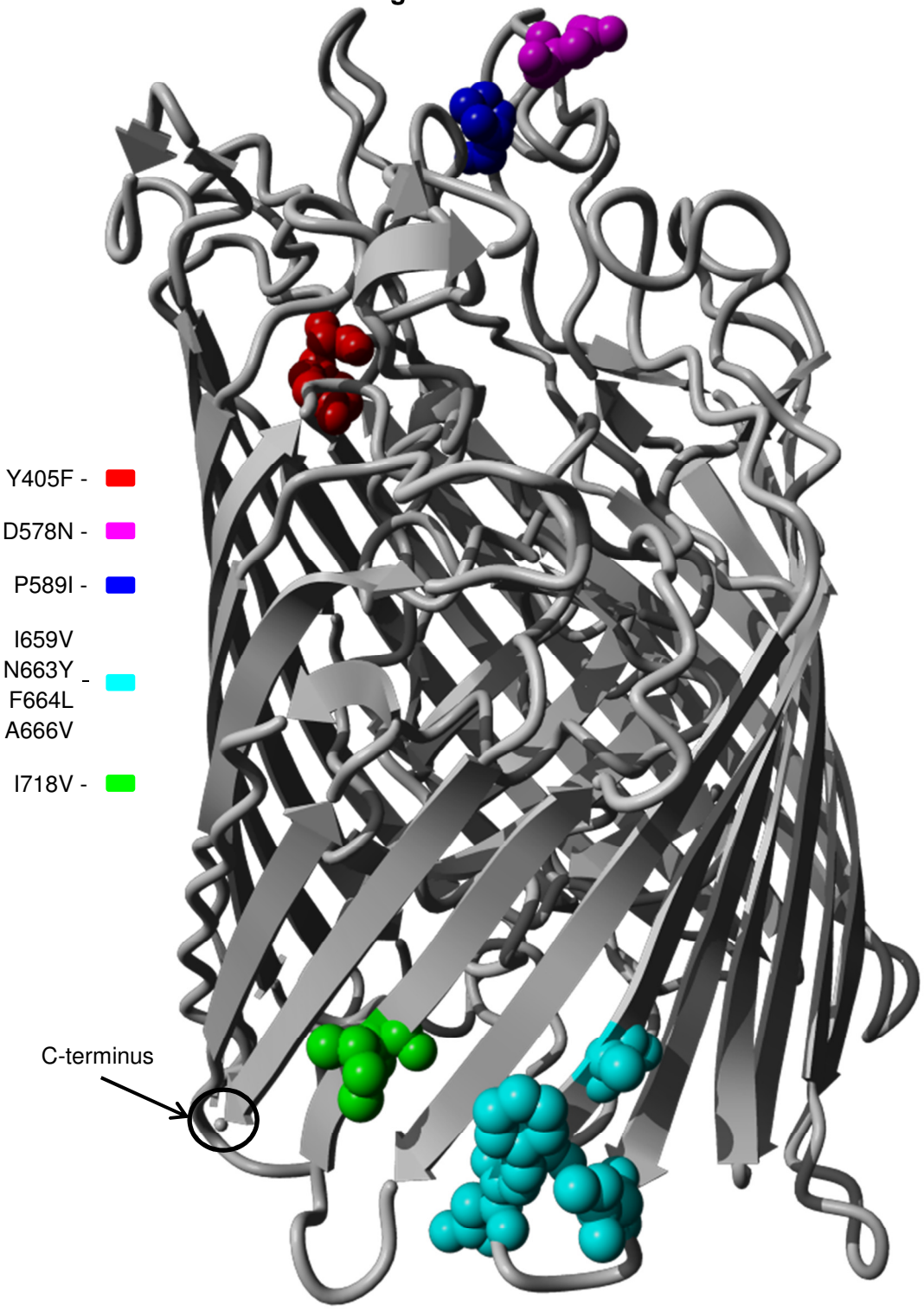


View from the periplasmic side of the predicted structure of HemR from ST57 isolate F199-3 showing the locations of polymorphisms A70V through T324I/E. The N-terminal signal peptide has been removed in this figure. The polymorphisms from Figure 4.8 are color coded as shown and have been depicted using a space-filling model for added visibility.

consensus amino acid, and if the PR < 1 the alternate amino acid is present).

It should be noted that the N-terminal 24 residue signal peptide identified in **Figure 4.1** has been removed in **Figures 4.6 – 4.8**. This region contains the first four polymorphisms (L4I/F, R8H, L15F, and V16I) in **Table 4.4**. It is unknown what effect, if any, alteration in this region may produce. Studies of

Figure 4.8



Side view of the predicted structure of HemR from ST57 isolate F199-3 showing the locations of polymorphisms Y405F through I718V. The N-terminal signal peptide has been removed in this figure. The five polymorphisms are color coded as shown and have been depicted using a space-filling model for added visibility.

human membrane and secreted proteins have identified mutations within signal peptide regions that are implicated in various diseases, including Wolman disease, Aspartylglucosaminuria, and Bernard–Soulier syndrome (54). It is thought that alterations within signal peptide domains impair the proper translocation and/or secretion of the proteins associated with these outcomes. The polymorphisms identified within the HemR signal peptide domain may also have a similar adverse effect on the proper localization of the protein.

It is apparent from **Figures 4.6 – 4.8** that the amino acid polymorphisms listed in **Table 4.4** occupy a variety of structural, and potentially functional, domains. Aside from the four polymorphisms within the signal peptide region listed above, four further alterations, A70V, R131K, K157R, and Q164K, are located in the plug domain that occludes the interior channel (**Figures 4.6 & 4.7**). In other, more widely researched TonB dependent receptors (primarily in *E. coli*), the plug domain is thought to either undergo a conformational change or exit the protein channel entirely upon interaction with TonB, thus allowing transit of the ligand (63). The four HemR polymorphisms identified within its plug domain could potentially alter these functions. Furthermore, in at least some TonB dependent receptors, regions on the extracellular side of the plug domain are involved in ligand recognition and binding (62). The A70V polymorphism (colored red in **Figures 4.6 & 4.7**) is in close proximity to the extracellular apex of the plug domain, and could theoretically influence ligand interaction.

Four other HemR polymorphisms, N303H/R and T324I/E (colored light blue and green in **Figures 4.6 & 4.7**, respectively) and I718V and the quartet of

polymorphisms beginning at I659V (colored green and light blue in **Figure 4.8**, respectively), are located within various β-strands that form the transmembrane domain. All but T324I/E are on the periplasmic side of the protein close to, but not within, intracellular loops. Interestingly, the I659V quartet surrounds one of the intracellular loops and includes both residues immediately adjacent, while the loop itself (consisting of three residues) is identical in all 146 NTHi isolates examined. The final three polymorphisms, Y405F, D578N, and P589I (**Figure 4.8**), are all positioned within extracellular loops, with D578N and P589I located within the same loop. The size of this loop is variable due to several small insertions and deletions, and D578N and P589I are separated by seven to ten amino acids. In *E. coli* and other bacteria, the extracellular loops of TonB dependent receptors are involved in ligand recognition and binding (9, 10, 15, 62, 63, 101, 132). The extracellular loops of HemR in NTHi likely have a similar role, and alterations within them could affect this function.

Adjustment for Population Structure

To adjust for the possible confounding effect of population structure on the association between HemR polymorphisms and otitis media, multiple logistic regression models were constructed in R version 2.13.0 (111) with the individual membership proportions for each population (as described in Chapter 3) coded as continuous variables (eight in total). A separate set of models was run for each polymorphism, and permutation tests were used to calculate p-values. In addition to the crude (i.e. unadjusted) models, models controlling for all eight

populations simultaneously as well as models adjusting for those populations that minimized the Akaike information criterion (AIC) values were created. The AIC is a measure of the relative goodness of fit of a statistical model, and offers a measure of the information lost when a given model is used to describe reality. Additionally, it includes a penalty that increases with the number of estimated parameters, thus discouraging overfitting. For a set of candidate models, the model with the minimum AIC value is preferred. Typically, models with AIC values within one or two of the minimum have substantial support, models with values within four to seven of the minimum have considerably less support, and models with values greater than ten above the minimum have essentially no support (12).

Table 4.5 presents the OR, permutation test p-value, and AIC value for all three sets of models run for each polymorphism. The unadjusted results are repeated from **Table 4.4** for comparison purposes. The mean AIC for the unadjusted models was 197.13, with a high of 199.26 for D578N and a low of 191.17 for A70V. Upon adjustment for all eight populations identified in Chapter 3, the majority (ten of fifteen) of the HemR polymorphism – otitis media associations lost statistical significance. Four of those remaining (R8H, A70V, Y405F and the I659V quartet) retained significance at the 95% confidence level, and D578N was significant at the 90% confidence level. The average AIC for these models was considerably lower (188.78) than the AIC values for the unadjusted models, indicating that the models adjusted for all populations were a better fit. Every polymorphism – disease association adjusted for all populations

Table 4.5

	Unadjusted			All Populations[a]			AIC Populations[b]		
	OR	p-value	AIC	OR	p-value	AIC	OR	p-value	AIC
L4I/F	3.52	0.02	197.11	1.50	0.26	189.54	1.35	0.31	183.30
R8H	0.43	0.04	198.74	4.50	0.00	187.36	5.01	0.00	180.47
L15F	4.03	0.00	192.69	1.74	0.14	188.99	1.93	0.09	181.79
V16I	3.14	0.03	198.23	2.00	0.14	188.79	2.25	0.10	181.54
A70V	0.13	0.00	191.17	0.37	0.05	189.31	0.11	0.00	180.95
R131K	3.97	0.02	196.86	1.42	0.30	189.62	1.43	0.29	183.25
K157R	3.14	0.03	198.23	1.01	0.50	189.80	1.11	0.43	183.49
Q164K	3.54	0.06	198.95	0.50	0.21	189.34	0.74	0.36	183.41
N303H/R	2.90	0.03	198.33	1.15	0.41	189.76	1.10	0.43	183.49
T324I/E	3.06	0.07	199.15	1.19	0.41	189.76	0.89	0.44	183.49
Y405F	5.38	0.02	197.27	5.34	0.03	186.28	3.92	0.05	180.29
D578N	2.77	0.07	199.26	2.42	0.09	187.67	2.37	0.09	181.34
P589I	2.09	0.02	197.73	1.60	0.12	189.04	1.02	0.49	183.51
I659V N663Y F664L A666V	2.58	0.01	195.78	2.27	0.03	187.07	1.75	0.08	181.52
I718V	2.18	0.02	197.44	1.40	0.21	189.32	1.46	0.16	182.60
	Average AIC		197.13			188.78			182.30

Association between HemR polymorphisms and otitis media, adjusted for population structure. All p-values were calculated by permutation tests, permuting the polymorphism data labels.

[a] Adjustment for all eight populations. Associations significant at the 90% confidence level are shaded blue.

[b] Adjustment for the populations that minimized the AIC. These were populations 2 and 8 for A70V, and populations 2, 7, and 8 for all other polymorphisms. Associations significant at the 90% confidence level are shaded green.

had a lower AIC than its unadjusted counterpart, though the AIC for A70V was only 1.86 lower.

Models with various combinations of populations as covariates were run until the model with the minimum AIC was found. With only a single exception, adjusting for populations 2, 7, and 8 resulted in the lowest AIC value. For A70V, the model that minimized the AIC was adjusted for populations 2 and 8. The strong collinearity between that polymorphism and population 7 may explain the difference (18 of 20 isolates with a valine at position 70 have greater than 99% of their ancestry from population 7, and one further claims a third of its ancestry from population 7). The five polymorphisms that were statistically significant at the 90% confidence level after adjusting for all population remained so, and a further two (L15F and V16I) joined their ranks. The average AIC for this set of models was 182.30, which was 6.48 lower than the mean AIC when adjusting for all populations and 14.83 lower than the mean unadjusted AIC. This suggests that controlling for all eight populations may have resulted in moderate overfitting, and that the lowest amount of information loss and the best fit was typically obtained by adjusting only for populations 2, 7, and 8.

The seven polymorphisms that remained statistically significant after adjustment for AIC minimizing population were not limited to particular structural or functional domains. R8H, L15F, and V16I are all located within the putative signal peptide, the A70V polymorphism is situated near the extracellular side of the plug domain, Y405F and D578N are positioned within different extracellular loops, and the four polymorphisms beginning with I659V are immediately

adjacent to a conserved intracellular loop. While population structure did have a confounding effect on the relationship between some HemR polymorphisms and otitis media, seven others remained statistically significant once that confounding was controlled for, once again implicating HemR, and various alleles thereof, in NTHi virulence.

DISCUSSION

Iron is an essential element for life in virtually all organisms, including *H. influenzae*. Hi has the additional absolute requirement of heme for aerobic growth, as it is unable to synthesize the heme precursor protoporphyrin IX (145). However, due to the instability of ferrous iron (the biologically relevant form) in the aerobic conditions and most pathogenic organisms' need for the element, nearly all iron in the human body is sequestered by high affinity iron binding proteins, such as transferrin, lactoferrin, and haptoglobin, or by incorporation within molecules such as heme (62, 63, 132). In response, bacteria have developed numerous, often redundant mechanisms for acquiring iron from their human hosts.

Nontypeable *H. influenzae* are no exception, and have a variable number of systems to scavenge iron and heme from many different host sources, including transferrin, hemoglobin, heme, heme:hemopexin, heme:albumin, and hemoglobin:haptoglobin complexes, and siderophores produced by other microorganisms (84, 144). A number of these systems have been implicated in Hi virulence. Using a chinchilla model of otitis media and 5- and 30-day-old rat

models of bacteremia, isogenic strains lacking certain iron acquisition genes have lower pathogenicity. These studies have found significant differences in virulence using the hemoglobin:haptoglobin binding proteins (*hgpA-C*) (79), the heme binding lipoprotein (*hbpA*) (81), lipoprotein *e* (P4) (*hel*) (83), and the heme:hemopexin utilization proteins (*hxuCBA*). Some iron acquisition systems have also been found at different frequencies in OM and commensal NTHi isolates. For example, in a 2006 study Xie et al. found by dot blot hybridization that *hgpB* was 1.36 times more prevalent among OM NTHi isolates than among commensal NTHi isolates (146). Together, these data present a convincing argument that the presence of genes for the acquisition of iron and heme from a variety of sources is important not just for growth in *H. influenzae*, but for pathogenesis as well.

However, many iron acquisition genes are present in the majority of NTHi strains assayed and present only modest prevalence differences between disease-causing and commensal isolates, if there is a significant difference at all. Such genes may still play an important role in pathogenesis, and certain alleles may be able to better provide NTHi with iron and heme in normally privileged environments such as the middle ear, thus enhancing virulence. To investigate this hypothesis, the full coding sequence for the hemin receptor *hemR* was obtained for 85 OM and 61 commensal NTHi isolates. Amino acid polymorphisms whose prevalences differed between the OM and commensal isolates were identified from the translated *hemR* nucleotide sequences and related to the theoretical three dimensional structure of the protein. Additionally,

the potentially confounding effect of the population structure characterized in the preceding chapter was controlled for in the analyses.

A total of 47 unique HemR amino acid sequences (or 'types') were identified among the 146 NTHi isolates used in the study. Most isolates with identical STs also had identical HemR sequences, and the majority of STs that shared a HemR sequence had a large proportion of their ancestry from the same population (as identified in Chapter 3). This is suggestive of two scenarios. In the simplest, the isolates within a population are more related and thus are more likely to have identical sequences at other loci. However, this fails to explain why isolates within HemR type 4, for example, have a wide range of ancestry proportions from population 2, from over 99% to just 50% (**Table 4.3**). Some of these isolates are, in fact, not very closely related (on an intraspecific scale, at any rate) despite having identical HemR sequences. Furthermore, five isolates of two STs that share HemR type 4 have the majority of their ancestry from another population altogether. Similar patterns are also evident in HemR type 8 of population 3, HemR type 12 of population 5, and HemR type 1 of population 8.

In the other scenario, there are a number of HemR sequences present within the NTHi populations, and recombination leads to both the admixture observed in the population structure as well as different STs sharing the same HemR type. Recombination may be more frequent among members of the same population, so most STs that share a HemR type have the majority of their ancestry from the same population. This hypothesis would also explain why multiple isolates of the same ST occasionally have different HemR types.

Further support for this hypothesis can be found in those STs that have a different primary population from the other STs with that HemR type (shaded in brown in **Table 4.3**). With only a single exception, these STs acquired the second highest proportion of their ancestry from the most common primary population for that HemR type. For example, in HemR type 4 the majority of isolates have the highest membership proportion in population 2. There are five isolates in that same HemR type whose primary population is 5, but they also have significant membership in population 2. This could be the result of one or more recombination events that brought in sizeable amounts of genetic material from population 2, including *hemR*. The truth, most likely, is a combination of the two hypotheses presented above.

Despite low amino acid sequence identity, the predicted structure of the NTHi HemR protein was remarkably similar to the experimentally derived structures of TonB dependent receptors involved in iron acquisition from other gram negative bacteria. A beta-barrel consisting of 22 anti-parallel β-strands composes a transmembrane channel that is occluded by an approximately 150 residue N-terminal plug domain, preventing the free diffusion of small molecules through the channel. In other systems, TonB interacts preferentially with ligand-bound receptors, inducing conformational changes in the plug domain that allow passage of the ligand through the receptor channel and into the periplasm. Even in these well studied systems, exactly how interaction between TonB and ligand-bound receptors results in translocation of the ligand across the outer membrane is unclear (63). However it works, it is likely that NTHi HemR operates in a

similar fashion, with TonB harnessing the proton motive force of the cytoplasmic membrane to provide the energy necessary for hemin translocation.

Prior to adjusting for population structure, 18 amino acid polymorphisms were found to have statistically significant prevalence differences between the OM and commensal isolates. These polymorphisms were not limited to a particular structural or functional domain, but were instead dispersed throughout the HemR sequence and were found in the putative signal peptide, the N-terminal plug domain, transmembrane β-strands, and extracellular loops. After the confounding effect of population structure was adjusted for, over half of the HemR polymorphism – otitis media associations identified during the crude analysis lost statistical significance. This is clear evidence that the assessment and control of population structure is just as important for association studies in bacteria as it is in human studies.

As mentioned previously, four of the polymorphisms (I659V, N663Y, F664L, and A666V) are very closely spaced and always co-occur. This may be the result of a common lineage between the isolates that contain the polymorphisms, but these isolates derive a wide range of their ancestry (from greater than 90% to less than 50%) from at least five of the eight identified populations, suggesting little commonality among them. It seems more likely that this group of polymorphisms has spread via recombination (alternatively, the consensus amino acids could be the relative newcomers that have proliferated). It is also possible that the co-occurrence of these polymorphisms is required for the proper folding or function of HemR; thus, if one changes, all must change for

the receptor to operate adequately. If this is the case, it is conceivable that isolates with only a subset of the four polymorphisms have a low or non-functioning HemR and are more likely to lose the gene entirely as the cost of producing the protein would not be offset by a beneficial gain in hemin acquisition. This would leave primarily isolates with either all four polymorphisms or none of them, as was observed in this study.

There are only two polymorphisms in which the alternate amino acid is more common among OM isolates than among commensal isolates, R8H and A70V, exposing a strong trend toward the consensus amino acids among NTHi isolates associated with otitis media. Interestingly, both polymorphisms seem to be associated with population 7, which is composed solely of OM isolates of ST57. Of the 30 isolates with histidine (the alternative amino acid) at position eight, 28 of them, including the 18 ST57 isolates, have at least a quarter of their ancestry from population 7. Twenty total isolates have a valine at position 70, of which 18 are the ST57 isolates and one further has a third of its ancestry from population 7. It is possible that the statistical significance of the association between otitis media and these two polymorphisms is the result of having multiple genotypically identical isolates within the study, analogous to the identification of the ST57 isolates as a distinct subpopulation in Chapter 3. Even if that is the case, however, the R8H and A70V polymorphisms may still be involved in virulence. As discussed in Chapter 3, all 18 ST57 isolates were collected from the middle ears of children with otitis media from all three geographic regions in roughly equal proportions. Despite being by far the most

commonly identified ST among the 170 NTHi isolates, it was completely unrepresented in the commensal isolates. This would appear to be strong evidence for heightened virulence in this genotype, and it is possible that these two HemR polymorphisms play a role, possibly by better providing heme to the bacteria while in the middle ear.

While a number of HemR amino acid polymorphisms were found to be significantly associated with otitis media after adjusting for population structure, no host or environmental factors were considered in the analyses, an important limitation of this study. As mentioned in Chapter 1, a number of host and environmental factors have been implicated in NTHi pathogenesis, including Eustachian tube dysfunction, preceding viral respiratory infection, allergies, exposure to cigarette smoke, and attending a daycare center (5, 6, 46). These factors, and others as yet undescribed, could influence the way bacterial virulence factors like HemR affect pathogenesis. Most association studies on bacterial virulence factors do not consider host or environmental factors, frequently because such information is simply not available, as is the case here. The bacterial strains utilized in such analyses are often originally isolated for other purposes and host and environment information is simply not collected. In other cases, privacy concerns or the cost associated with collecting such information can be significant barriers. Nevertheless, association studies linking bacterial factors to disease have yielded a great deal of invaluable knowledge, and provide an effective starting point for future studies to incorporate the effects of bacterial, host, and environmental factors into a cohesive whole.

Chapter 5

Conclusion and Future Directions

Nontypeable *Haemophilus influenzae* are small, gram negative bacteria whose only natural hosts are humans. Asymptomatic colonization of the naso- and oropharynges is common, particularly in young children, and there is frequently simultaneous colonization with multiple strains. NTHi can also cause a range of respiratory diseases, chiefly otitis media. The mechanisms and determinants of virulence in NTHi have been the focus of considerable research, and the presence of a number of genetic loci has been associated with otitis media, including adhesins, outer membrane proteins, and genes involved in iron acquisition and lipooligosaccharide biosynthesis. Other subjects, such as the population structure of NTHi and the involvement in virulence of genes present in both commensal and OM isolates, are less well studied.

The research presented in the previous three chapters addresses these issues in NTHi, though, of course, much work remains. Chapter 2 used multilocus sequence typing to explore the diversity of the NTHi serially isolated from two healthy children attending a daycare center. Despite previous identification as NTHi using the then standard laboratory methods, 25 of the 46 total isolates were found to be non-NTHi using more sensitive techniques. Interestingly, 12 of the 25 non-NTHi isolates reacted with an ostensibly Hi

specific antibody but clustered with the closely related but non-pathogenic *H. haemolyticus*, suggesting that these isolates may represent strains that are intermediate between Hi and *H. haemolyticus* on the genetic continuum between species. Of the nine STs identified among the 21 NTHi isolates from the two children, most were only found during a single sampling period, and no ST was shared between the two children despite concurrent attendance at the same daycare center. Furthermore, eBURST and maximum parsimony analysis revealed little clustering among the isolates, indicating they are not close relatives on an intraspecific scale. These data highlight both the high level of genetic diversity present among NTHi colonizing healthy children, as well as the inadequacy of the bacterial isolation technique used during collection to detect all unique STs present during a single sampling period.

In Chapter 3, diversity and population structure was assessed in a larger collection of 170 NTHi isolates, both OM and commensal, from three geographic regions. Using MLST, 109 unique STs were identified, 53 of which were previously undescribed. Interestingly, 11 of these STs had a complete deletion of the fucose operon, which contains the MLST loci *fucK*. This group of STs, while still within our laboratory's definition of NTHi, appears to occupy an intermediate position between the more typical NTHi and *H. haemolyticus*, again showing that the often arbitrary delineations between bacterial species are inadequate in describing the true relationships between bacteria. Evaluation of population structure identified support for eight populations when all isolates were included in the analysis. Six of these populations were also identified when

only unique genotypes were analyzed, while each of the other two (populations 7 and 8) consisted of multiple genotypically identical isolates. These two populations, along with the *fucK* negative isolates (which formed population 2), are of interest for studies of NTHi virulence, though for opposite reasons. Populations 7 and 8 consist solely of OM-associated isolates of ST57 and ST34, respectively, while population 2 is composed of only commensal isolates missing *fucK*. Investigating the differences between these populations could reveal important information on the mechanisms of NTHi virulence.

Chapter 4 examined the full coding sequence of the hemin receptor *hemR* to identify amino acid polymorphisms associated with otitis media in 146 of the NTHi isolates characterized in Chapter 3. Because the presence of population structure can significantly confound studies of association, each HemR polymorphism-otitis media association was adjusted for the population structure identified in Chapter 3. Additionally, the predicted three dimensional protein structure of HemR was investigated, and potential structural and functional domains were defined. A total of 47 unique HemR amino acid sequences were identified among the 146 isolates, and in the crude analyses, 18 polymorphisms were found to have statistically significant prevalence differences between OM and commensal isolates. After adjusting for population structure, over half of these polymorphisms lost statistical significance, demonstrating the biasing effect of population structure and the importance of taking it into account in studies of association. However, seven polymorphisms retained significance (counting the quartet of polymorphisms starting at I659V as a single entity), implicating HemR

in NTHi virulence and suggesting that certain alleles may be better able to supply the bacteria with hemin.

The 170 isolates utilized in Chapters 3 and 4 represent a valuable resource for further studies on a variety of topics relevant to NTHi. They have been randomly selected from different children in three geographic regions, their identity as NTHi has been confirmed by phylogeny, they have been well characterized by multilocus sequence typing, and population structure within the sample has been thoroughly assessed. This will provide an expedient backdrop against which other studies may be undertaken. For example, the prevalence of other iron acquisition genes (or any potential virulence gene) among OM and commensal isolates could be assessed, as well as their relationship with the population structure and geographic area. Similarly, individual polymorphisms and alleles of said genes could be studied for associations with virulence, population structure, or geography. It would also be interesting to examine TonB in this collection, as it interacts with many high affinity outer membrane receptors, including those for iron and iron containing molecules.

One avenue of further study is in identifying functional differences associated with the HemR polymorphisms identified in Chapter 4. Creating isogenic mutants that vary only in their *hemR* sequence is likely to be the best scheme, as this would eliminate other, potentially confounding variables. One method to identify differences would be to assess the growth of the isogenic strains with various polymorphisms, both singly and in combination, on media with different levels and types of iron. This would determine if there are any

basic deficiencies or advantages inherent to the different HemR polymorphisms.

A complementary method would be to investigate the comparative virulence of these strains in the chinchilla model of otitis media. Chinchillas are the model of choice for otitis media as their cochlea are comparatively easy to access and of a relatively similar size to that of humans, and the disease can be produced by a small inocula injected into the middle ear (41). This model system has been extensively used to assess the *in vivo* effect of various NTHi virulence factors, including the hemoglobin-haptoglobin binding proteins HgpA-D and the heme-binding lipoprotein HbpA (73, 79, 81, 100). Assessing the comparative virulence of isogenic NTHi strains with the HemR amino acid polymorphisms identified in this study, both singly and in combination, within this model system would provide a more complete picture of their potential effects on virulence in their human hosts.

There are a number of potential issues with studies in animal models, of course. The issue likely to be the largest obstacle is the limitations inherent to using model systems. This is particularly true of human specific pathogens like Hi, where the bacteria do not normally colonize the animal model and the experimental modes of infection may not resemble those found in nature. Growth on media and in animal models is at best a poor representation of the reality of NTHi colonization and disease, and it is difficult to determine how generalizable a significant result, or lack thereof, found in chinchillas is to a human child. Furthermore, as discussed previously in Chapters 1 and 4, *H. influenzae* has a number of at least partially overlapping systems for the uptake

of iron, heme, and related molecules. This redundancy can make it difficult to find true phenotypic differences, as other iron uptake systems may be able to compensate under the artificial conditions of model systems. Nevertheless, animal models have significant advantages over *in vitro* systems, primarily in the form of intact, organized organ and immune systems, which can greatly aid our ability to characterize the pathogenesis of and immune response to microbial infection.

An examination of the host and environmental factors that impact NTHi virulence, and how they interact with bacterial virulence factors, would be extremely interesting as well. As mentioned in Chapters 1 and 4, studies have identified a number of host and environmental factors that are associated with colonization and disease, and others have identified bacterial factors that are similarly associated, but few studies have attempted to examine the intersection between all three in a cohesive fashion. Such studies can be extraordinarily difficult and resource intensive, as the relationship between host, environment, and pathogen is typically complex and massive amounts of data must be collected to accurately and coherently piece it together. Most likely, a new set of isolates would need to be collected from a well-defined population in order to obtain the necessary information, as most extant collections lack at least some information regarding host and environmental factors. Of particular interest would be how host immune processes interact with environmental and bacterial factors during otitis media. Given the complexity of the human immune system, however, having pre-determined targets for study would likely be a necessity.

The chinchilla model of otitis media, while far from perfect, could once again play a useful role here, with which it would be far easier and less costly to identify potentially significant immune system components. These host factors could then be targeted for study in human otitis media and the way in which they affect, and are affected by, bacterial and environmental factors could be assessed.

Finally, a thorough characterization of the 18 ST57 isolates that comprised population 7 would be of considerable interest. As discussed previously, ST57 was the most commonly identified ST among the 170 NTHi isolates from Chapter 3. Sequence type 57 isolates were found in approximately equal proportions from all three geographic regions, but only from the middle ears of children with otitis media. As no two ST57 isolates were collected from the same child, this suggests a strong association between that genotype and otitis media, but no association with geography. Furthermore, despite the naso- and oropharynges being the reservoir for NTHi, ST57 was not identified among the commensal isolates. These isolates may represent a rare genotype with relatively high virulence, leading to its increased presence among disease-associated isolates. Determining what causes this increased pathogenicity would be extremely interesting. One potential method would be to utilize a technique like microarray hybridization to identify differences in genome content between ST57 isolates and other, less pathogenic isolates (such as the *fucK* negative population 2 isolates). Furthermore, the rapidly decreasing costs of DNA sequencing makes obtaining the full genome sequence for at least one ST57 isolate a realistic proposition. Acquiring full genome sequences for several ST57 isolates (one

isolate from each of the three geographic regions, for example) would make comparisons between isolates possible, and determining exactly how similar these isolates are to each other would be quite interesting. Comparisons of genomic content and sequence between ST57 isolates and other fully sequences isolates, particularly the commensal NTHi strains in the assembly stage of the NCBI genome project webpage, could reveal fascinating differences that help illuminate the mechanisms of virulence in NTHi.

References

1. **Achtman, M.** 2004. Population structure of pathogenic bacteria revisited. Int J Med Microbiol **294:**67-73.

2. **American Academy of Pediatrics Subcommittee on Management of Acute Otitis Media.** 2004. Diagnosis and management of acute otitis media. Pediatrics **113:**1451-65.

3. **Aracil, B., M. Slack, M. Perez-Vazquez, F. Roman, M. Ramsay, and J. Campos.** 2006. Molecular epidemiology of Haemophilus influenzae type b causing vaccine failures in the United Kingdom. J Clin Microbiol **44:**1645-9.

4. **Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak.** 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol **340:**783-95.

5. **Berman, S.** 1995. Otitis media in children. N Engl J Med **332:**1560-5.

6. **Bhetwal, N., and J. R. McConaghy.** 2007. The evaluation and treatment of children with acute otitis media. Prim Care **34:**59-70.

7. **Block, S. L., J. Hedrick, C. J. Harrison, R. Tyler, A. Smith, R. Findlay, and E. Keegan.** 2004. Community-wide vaccination with the heptavalent pneumococcal conjugate significantly alters the microbiology of acute otitis media. Pediatr Infect Dis J **23:**829-33.

8. **Bou, R., A. Dominguez, D. Fontanals, I. Sanfeliu, I. Pons, J. Renau, V. Pineda, E. Lobera, C. Latorre, M. Majo, and L. Salleras.** 2000. Prevalence of *Haemophilus influenzae* pharyngeal carriers in the school population of Catalonia. Working Group on invasive disease caused by *Haemophilus influenzae*. Eur J Epidemiol **16:**521-6.

9. **Braun, V., and F. Endriss.** 2007. Energy-coupled outer membrane transport proteins and regulatory proteins. Biometals **20:**219-31.

10. **Buchanan, S. K., B. S. Smith, L. Venkatramani, D. Xia, L. Esser, M. Palnitkar, R. Chakraborty, D. van der Helm, and J. Deisenhofer.** 1999.

Crystal structure of the outer membrane active transporter FepA from Escherichia coli. Nat Struct Biol **6:**56-63.

11. **Budroni, S., E. Siena, J. C. Hotopp, K. L. Seib, D. Serruto, C. Nofroni, M. Comanducci, D. R. Riley, S. C. Daugherty, S. V. Angiuoli, A. Covacci, M. Pizza, R. Rappuoli, E. R. Moxon, H. Tettelin, and D. Medini.** 2011. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A **108:**4494-9.

12. **Burnham, K. P., and D. R. Anderson.** 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. Springer, New York.

13. **Caniato, F. F., C. T. Guimaraes, M. Hamblin, C. Billot, J. F. Rami, B. Hufnagel, L. V. Kochian, J. Liu, A. A. Garcia, C. T. Hash, P. Ramu, S. Mitchell, S. Kresovich, A. C. Oliveira, G. de Avellar, A. Borem, J. C. Glaszmann, R. E. Schaffert, and J. V. Magalhaes.** 2011. The Relationship between Population Structure and Aluminum Tolerance in Cultivated Sorghum. PLoS ONE **6:**e20830.

14. **Casey, J. R., and M. E. Pichichero.** 2004. Changes in frequency and pathogens causing acute otitis media in 1995-2003. Pediatr Infect Dis J **23:**824-8.

15. **Chakraborty, R., E. A. Lemke, Z. Cao, P. E. Klebba, and D. van der Helm.** 2003. Identification and mutational studies of conserved amino acids in the outer membrane receptor protein, FepA, which affect transport but not binding of ferric-enterobactin in Escherichia coli. Biometals **16:**507-18.

16. **Clemans, D. L., C. F. Marrs, M. Patel, M. Duncan, and J. R. Gilsdorf.** 1998. Comparative analysis of *Haemophilus influenzae hifA* (pilin) genes. Infect Immun **66:**656-63.

17. **Cody, A. J., D. Field, E. J. Feil, S. Stringer, M. E. Deadman, A. G. Tsolaki, B. Gratz, V. Bouchet, R. Goldstein, D. W. Hood, and E. R. Moxon.** 2003. High rates of recombination in otitis media isolates of non-typeable *Haemophilus influenzae*. Infect Genet Evol **3:**57-66.

18. **Davis, G. S., S. A. Sandstedt, M. Patel, C. F. Marrs, and J. R. Gilsdorf.** 2011. Use of bexB to detect the capsule locus in Haemophilus influenzae. J Clin Microbiol.

19. **Davis, J., A. L. Smith, W. R. Hughes, and M. Golomb.** 2001. Evolution of an autotransporter: domain shuffling and lateral transfer from pathogenic *Haemophilus* to *Neisseria*. J Bacteriol **183:**4626-35.

20. **Dean, A. G., K. M. Sullivan, and M. M. Soe** 2010/09/19, posting date. OpenEpi: open source epidemiologic statistics for public health, version 2.3.1. [Online.]

21. **den Bakker, H. C., X. Didelot, E. D. Fortes, K. K. Nightingale, and M. Wiedmann.** 2008. Lineage specific recombination rates and microevolution in Listeria monocytogenes. BMC evolutionary biology **8:**277.

22. **Dhooge, I., M. Vaneechoutte, G. Claeys, G. Verschraegen, and P. Van Cauwenberge.** 2000. Turnover of *Haemophilus influenzae* isolates in otitis-prone children. Int J Pediatr Otorhinolaryngol **54:**7-12.

23. **Didelot, X., and D. Falush.** 2008. ClonalFrame User Guide.

24. **Didelot, X., and D. Falush.** 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics **175:**1251-66.

25. **Ecevit, I. Z., K. W. McCrea, M. M. Pettigrew, A. Sen, C. F. Marrs, and J. R. Gilsdorf.** 2004. Prevalence of the *hifBC*, *hmw1A*, *hmw2A*, *hmwC*, and *hia* Genes in *Haemophilus influenzae* Isolates. J Clin Microbiol **42:**3065-72.

26. **Erwin, A. L., K. L. Nelson, T. Mhlanga-Mutangadura, P. J. Bonthuis, J. L. Geelhood, G. Morlin, W. C. Unrath, J. Campos, D. W. Crook, M. M. Farley, F. W. Henderson, R. F. Jacobs, K. Muhlemann, S. W. Satola, L. van Alphen, M. Golomb, and A. L. Smith.** 2005. Characterization of genetic and phenotypic diversity of invasive nontypeable *Haemophilus influenzae*. Infect Immun **73:**5853-63.

27. **Erwin, A. L., S. A. Sandstedt, P. J. Bonthuis, J. L. Geelhood, K. L. Nelson, W. C. Unrath, M. A. Diggle, M. J. Theodore, C. R. Pleatman, E.**

A. Mothershed, C. T. Sacchi, L. W. Mayer, J. R. Gilsdorf, and A. L. Smith. 2008. Analysis of genetic relatedness of *Haemophilus influenzae* isolates by multilocus sequence typing. J Bacteriol **190**.

28. **Erwin, A. L., and A. L. Smith.** 2007. Nontypeable *Haemophilus influenzae*: understanding virulence and commensal behavior. Trends Microbiol **15:**355-62.

29. **Eskola, J., and T. Kilpi.** 2000. Potential of bacterial vaccines in the prevention of acute otitis media. Pediatr Infect Dis J **19:**S72-8.

30. **Eskola, J., T. Kilpi, A. Palmu, J. Jokinen, J. Haapakoski, E. Herva, A. Takala, H. Kayhty, P. Karma, R. Kohberger, G. Siber, and P. H. Makela.** 2001. Efficacy of a pneumococcal conjugate vaccine against acute otitis media. N Engl J Med **344:**403-9.

31. **Falla, T. J., D. W. Crook, L. N. Brophy, D. Maskell, J. S. Kroll, and E. R. Moxon.** 1994. PCR for capsular typing of Haemophilus influenzae. J Clin Microbiol **32:**2382-6.

32. **Falush, D., and R. Bowden.** 2006. Genome-wide association mapping in bacteria? Trends Microbiol **14:**353-5.

33. **Falush, D., M. Stephens, and J. K. Pritchard.** 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164:**1567-87.

34. **Falush, D., T. Wirth, B. Linz, J. K. Pritchard, M. Stephens, M. Kidd, M. J. Blaser, D. Y. Graham, S. Vacher, G. I. Perez-Perez, Y. Yamaoka, F. Megraud, K. Otto, U. Reichard, E. Katzowitsch, X. Wang, M. Achtman, and S. Suerbaum.** 2003. Traces of human migrations in *Helicobacter pylori* populations. Science **299:**1582-5.

35. **Farjo, R. S., B. Foxman, M. J. Patel, L. Zhang, M. M. Pettigrew, S. I. McCoy, C. F. Marrs, and J. R. Gilsdorf.** 2004. Diversity and sharing of *Haemophilus influenzae* strains colonizing healthy children attending day-care centers. Pediatr Infect Dis J **23:**41-6.

36. **Feil, E. J.** 2004. Small change: keeping pace with microevolution. Nat Rev Microbiol **2:**483-95.

37. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol **186:**1518-30.

38. **Fernaays, M. M., A. J. Lesse, X. Cai, and T. F. Murphy.** 2006. Characterization of *igaB*, a second immunoglobulin A1 protease gene in nontypeable *Haemophilus influenzae*. Infect Immun **74:**5860-70.

39. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:**496-512.

40. **Gelman, A., and D. B. Rubin.** 1992. Inference from iterative simulation using multiple sequences. Stat Sci **7:**16.

41. **Giebink, G. S.** 1999. Otitis media: the chinchilla model. Microb Drug Resist **5:**57-72.

42. **Glasziou, P. P., C. B. Del Mar, S. L. Sanders, and M. Hayem.** 2004. Antibiotics for acute otitis media in children. Cochrane Database Syst Rev**:**CD000219.

43. **Goloboff, P., J. Farris, and K. Nixon.** 2003. T.N.T.: Tree Analysis Using New Technology. Program and documentation available fom the authors, and at www.zmuc.dk/public/phylogeny. Version 1.1.

44. **Greenberg, D., A. Broides, I. Blancovich, N. Peled, N. Givon-Lavi, and R. Dagan.** 2004. Relative importance of nasopharyngeal versus oropharyngeal sampling for isolation of Streptococcus pneumoniae and Haemophilus influenzae from healthy and sick individuals varies with age. J Clin Microbiol **42:**4604-9.

45. **Hanage, W. P., C. Fraser, and B. G. Spratt.** 2005. Fuzzy species among recombinogenic bacteria. BMC Biol **3:**6.

46. **Hardy, G. G., S. M. Tudor, and J. W. St Geme, 3rd.** 2003. The pathogenesis of disease due to nontypeable *Haemophilus influenzae*, p. 1-28. *In* M. A. Herbert, D. W. Hood, and E. R. Moxon (ed.), *Haemophilus influenzae* Protocols. Humana Press, Totowa, NJ.

47. **Harrison, A., D. W. Dyer, A. Gillaspy, W. C. Ray, R. Mungur, M. B. Carson, H. Zhong, J. Gipson, M. Gipson, L. S. Johnson, L. Lewis, L. O. Bakaletz, and R. S. Munson, Jr.** 2005. Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. J Bacteriol **187:**4627-36.

48. **Hoberman, A., J. L. Paradise, H. E. Rockette, N. Shaikh, E. R. Wald, D. H. Kearney, D. K. Colborn, M. Kurs-Lasky, S. Bhatnagar, M. A. Haralam, L. M. Zoffel, C. Jenkins, M. A. Pope, T. L. Balentine, and K. A. Barbadora.** 2011. Treatment of acute otitis media in children under 2 years of age. N Engl J Med **364:**105-15.

49. **Huson, D. H., and D. Bryant.** 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol **23:**254-67.

50. **Jacobs, M. M., M. J. Smulders, R. G. van den Berg, and B. Vosman.** 2011. What's in a name; genetic structure in Solanum section Petota studied using population-genetic tools. BMC Evol Biol **11:**42.

51. **Jakobsson, M., and N. A. Rosenberg.** 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics **23:**1801-6.

52. **Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H. C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton.** 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451:**998-1003.

53. **Janeway, C. A., P. Travers, M. Walport, and M. J. Shlomchik.** 2005. Immunobiology: The Immune System in Health and Disease, 6 ed. Garland Science, New York.

54. **Jarjanazi, H., S. Savas, N. Pabalan, J. W. Dennis, and H. Ozcelik.** 2008. Biological implications of SNPs in signal peptide domains of human proteins. Proteins **70:**394-403.

55. **Juliao, P. C., C. F. Marrs, J. Xie, and J. R. Gilsdorf.** 2007. Histidine auxotrophy in commensal and disease-causing nontypeable *Haemophilus influenzae*. J Bacteriol **189:**4994-5001.

56. **Kaplan, B., T. L. Wandstrat, and J. R. Cunningham.** 1997. Overall cost in the treatment of otitis media. Pediatr Infect Dis J **16:**S9-11.

57. **Kilian, M.** 2005. Genus *Haemophilus*, p. 883-904. *In* G. M. Garrity, D. J. Brenner, N. R. Krieg, and J. T. Staley (ed.), Bergey's Manual of Systematic Bacteriology, 2 ed, vol. 2. Springer-Verlag, New York.

58. **Kilpi, T., E. Herva, T. Kaijalainen, R. Syrjanen, and A. K. Takala.** 2001. Bacteriology of acute otitis media in a cohort of Finnish children followed for the first two years of life. Pediatr Infect Dis J **20:**654-62.

59. **Klein, J. O.** 2011. Is acute otitis media a treatable disease? N Engl J Med **364:**168-9.

60. **Kopelman, N. M., L. Stone, C. Wang, D. Gefel, M. W. Feldman, J. Hillel, and N. A. Rosenberg.** 2009. Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations. BMC Genet **10:**80.

61. **Krasan, G. P., D. Cutter, S. L. Block, and J. W. St Geme, 3rd.** 1999. Adhesin expression in matched nasopharyngeal and middle ear isolates of nontypeable Haemophilus influenzae from children with acute otitis media. Infect Immun **67:**449-54.

62. **Krewulak, K. D., and H. J. Vogel.** 2008. Structural biology of bacterial iron uptake. Biochim Biophys Acta **1778:**1781-804.

63. **Krewulak, K. D., and H. J. Vogel.** 2011. TonB or not TonB: is that the question? Biochem Cell Biol **89:**87-97.

64. **Krieger, E., G. Koraimann, and G. Vriend.** 2002. Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins **47:**393-402.

65. **LaCross, N. C., C. F. Marrs, M. Patel, S. A. Sandstedt, and J. R. Gilsdorf.** 2008. High Genetic Diversity of Nontypeable Haemophilus influenzae Among Two Children Attending a Daycare Center. J Clin Microbiol.

66. **Leduc, I., K. E. Banks, K. R. Fortney, K. B. Patterson, S. D. Billings, B. P. Katz, S. M. Spinola, and C. Elkins.** 2008. Evaluation of the repertoire of the TonB-dependent Receptors of *Haemophilus ducreyi* for their role in virulence in humans. J Infect Dis **197:**1103-9.

67. **Leibovitz, E., L. Piglansky, S. Raiz, D. Greenberg, K. A. Hamed, J. M. Ledeine, J. Press, A. Leiberman, R. M. Echols, P. F. Pierce, M. R. Jacobs, and R. Dagan.** 2003. Bacteriologic and clinical efficacy of oral gatifloxacin for the treatment of recurrent/nonresponsive acute otitis media: an open label, noncomparative, double tympanocentesis study. Pediatr Infect Dis J **22:**943-9.

68. **Little, P., C. Gould, I. Williamson, M. Moore, G. Warner, and J. Dunleavey.** 2001. Pragmatic randomised controlled trial of two prescribing strategies for childhood acute otitis media. BMJ **322:**336-42.

69. **Lund, M. E., and D. J. Blazevic.** 1977. Rapid speciation of *Haemophilus* with the porphyrin production test versus the satellite test for X. J Clin Microbiol **5:**142-4.

70. **Maiden, M. C.** 2006. Multilocus sequence typing of bacteria. Annu Rev Microbiol **60:**561-88.

71. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A **95:**3140-5.

72. **Martin, K., G. Morlin, A. Smith, A. Nordyke, A. Eisenstark, and M. Golomb.** 1998. The tryptophanase gene cluster of *Haemophilus influenzae* type b: evidence for horizontal gene transfer. J Bacteriol **180:**107-18.

73. **Mason, K. M., R. S. Munson, Jr., and L. O. Bakaletz.** 2003. Nontypeable *Haemophilus influenzae* gene expression induced in vivo in a chinchilla model of otitis media. Infect Immun **71:**3454-62.

74. **May, B. J., Q. Zhang, L. L. Li, M. L. Paustian, T. S. Whittam, and V. Kapur.** 2001. Complete genomic sequence of Pasteurella multocida, Pm70. Proc Natl Acad Sci U S A **98:**3460-5.

75. **McCrea, K. W., J. Xie, N. LaCross, M. Patel, D. Mukundan, T. F. Murphy, C. F. Marrs, and J. R. Gilsdorf.** 2008. Relationships of nontypeable *Haemophilus influenzae* strains to hemolytic and nonhemolytic *Haemophilus haemolyticus* strains. J Clin Microbiol **46:**406-16.

76. **McVean, G., P. Awadalla, and P. Fearnhead.** 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:**1231-41.

77. **Meats, E., E. J. Feil, S. Stringer, A. J. Cody, R. Goldstein, J. S. Kroll, T. Popovic, and B. G. Spratt.** 2003. Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J Clin Microbiol **41:**1623-36.

78. **Moor, P. E., P. C. Collignon, and G. L. Gilbert.** 1999. Pulsed-field gel electrophoresis used to investigate genetic diversity of Haemophilus influenzae type b isolates in Australia shows differences between Aboriginal and non-Aboriginal isolates. J Clin Microbiol **37:**1524-31.

79. **Morton, D. J., L. O. Bakaletz, J. A. Jurcisek, T. M. VanWagoner, T. W. Seale, P. W. Whitby, and T. L. Stull.** 2004. Reduced severity of middle ear infection caused by nontypeable *Haemophilus influenzae* lacking the hemoglobin/hemoglobin-haptoglobin binding proteins (Hgp) in a chinchilla model of otitis media. Microb Pathog **36:**25-33.

80. **Morton, D. J., L. L. Madore, A. Smith, T. M. Vanwagoner, T. W. Seale, P. W. Whitby, and T. L. Stull.** 2005. The heme-binding lipoprotein (HbpA) of *Haemophilus influenzae*: role in heme utilization. FEMS Microbiol Lett **253:**193-9.

81. **Morton, D. J., T. W. Seale, L. O. Bakaletz, J. A. Jurcisek, A. Smith, T. M. Vanwagoner, P. W. Whitby, and T. L. Stull.** 2009. The heme-binding protein (HbpA) of Haemophilus influenzae as a virulence determinant. Int J Med Microbiol.

82. **Morton, D. J., T. W. Seale, L. L. Madore, T. M. VanWagoner, P. W. Whitby, and T. L. Stull.** 2007. The haem-haemopexin utilization gene cluster (*hxuCBA*) as a virulence factor of *Haemophilus influenzae*. Microbiology **153:**215-24.

83. **Morton, D. J., A. Smith, T. M. VanWagoner, T. W. Seale, P. W. Whitby, and T. L. Stull.** 2007. Lipoprotein e (P4) *of Haemophilus influenzae*: role in heme utilization and pathogenesis. Microbes Infect **9:**932-9.

84. **Morton, D. J., E. J. Turman, P. D. Hensley, T. M. VanWagoner, T. W. Seale, P. W. Whitby, and T. L. Stull.** 2010. Identification of a siderophore utilization locus in nontypeable Haemophilus influenzae. BMC Microbiol **10:**113.

85. **Moxon, E. R., and K. A. Vaughn.** 1981. The type b capsular polysaccharide as a virulence determinant of *Haemophilus influenzae*: studies using clinical isolates and laboratory transformants. J Infect Dis **143:**517-24.

86. **Mukundan, D., Z. Ecevit, M. Patel, C. F. Marrs, and J. R. Gilsdorf.** 2007. Pharyngeal colonization dynamics of *Haemophilus influenzae* and *Haemophilus haemolyticus* in healthy adult carriers. J Clin Microbiol **45:**3207-17.

87. **Munson, R. S., Jr., A. Harrison, A. Gillaspy, W. C. Ray, M. Carson, D. Armbruster, J. Gipson, M. Gipson, L. Johnson, L. Lewis, D. W. Dyer, and L. O. Bakaletz.** 2004. Partial analysis of the genomes of two nontypeable *Haemophilus influenzae* otitis media isolates. Infect Immun **72:**3002-10.

88. **Murphy, T. F., A. L. Brauer, S. Sethi, M. Kilian, X. Cai, and A. J. Lesse.** 2007. *Haemophilus haemolyticus*: a human respiratory tract commensal to be distinguished from *Haemophilus influenzae*. J Infect Dis **195:**81-9.

89. **Murphy, T. F., C. Kirkham, and D. J. Sikkema.** 1992. Neonatal, urogenital isolates of biotype 4 nontypeable *Haemophilus influenzae* express a variant P6 outer membrane protein molecule. Infect Immun **60:**2016-22.

90. **Murphy, T. F., S. Sethi, K. L. Klingman, A. B. Brueggemann, and G. V. Doern.** 1999. Simultaneous respiratory tract colonization by multiple strains of nontypeable *Haemophilus influenzae* in chronic obstructive pulmonary disease: implications for antibiotic therapy. J Infect Dis **180:**404-9.

91. **Musser, J. M., S. J. Barenkamp, D. M. Granoff, and R. K. Selander.** 1986. Genetic relationships of serologically nontypable and serotype b strains of *Haemophilus influenzae*. Infect Immun **52:**183-91.

92. **Musser, J. M., D. M. Granoff, P. E. Pattison, and R. K. Selander.** 1985. A population genetic framework for the study of invasive diseases caused by serotype b strains of Haemophilus influenzae. Proceedings of the National Academy of Sciences of the United States of America **82:**5078-82.

93. **Musser, J. M., J. S. Kroll, D. M. Granoff, E. R. Moxon, B. R. Brodeur, J. Campos, H. Dabernat, W. Frederiksen, J. Hamel, G. Hammond, E. A. Hoiby, K. E. Jonsdottir, M. Kabeer, I. Kallings, H. J. Koornhof, B. Law, K. I. Li, J. Montgomery, P. E. Pattison, J. Piffaretti, A. K. Takala, M. L. Thong, R. A. Wall, J. I. Ward, and R. K. Selander.** 1990. Global genetic structure and molecular epidemiology of encapsulated *Haemophilus influenzae*. Rev Infect Dis **12:**75-111.

94. **Musser, J. M., J. S. Kroll, E. R. Moxon, and R. K. Selander.** 1988. Clonal population structure of encapsulated *Haemophilus influenzae*. Infect Immun **56:**1837-45.

95. **Musser, J. M., J. S. Kroll, E. R. Moxon, and R. K. Selander.** 1988. Evolutionary genetics of the encapsulated strains of Haemophilus influenzae. Proc Natl Acad Sci U S A **85:**7758-62.

96. **Nielsen, H., and A. Krogh.** 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol **6:**122-30.

97. **Norskov-Lauritsen, N.** 2009. Detection of cryptic genospecies misidentified as Haemophilus influenzae in routine clinical samples by assessment of marker genes fucK, hap, and sodC. J Clin Microbiol **47:**2590-2.

98. **Norskov-Lauritsen, N., B. Bruun, and M. Kilian.** 2005. Multilocus sequence phylogenetic study of the genus *Haemophilus* with description of *Haemophilus pittmaniae* sp. nov. Int J Syst Evol Microbiol **55:**449-56.

99. **Norskov-Lauritsen, N., M. D. Overballe, and M. Kilian.** 2009. Delineation of the species Haemophilus influenzae by phenotype, multilocus sequence phylogeny, and detection of marker genes. J Bacteriol **191:**822-31.

100. **Novotny, L. A., S. Partida-Sanchez, R. S. Munson, Jr., and L. O. Bakaletz.** 2008. Differential uptake and processing of a *Haemophilus influenzae* P5-derived immunogen by chinchilla dendritic cells. Infect Immun **76:**967-77.

101. **Pawelek, P. D., N. Croteau, C. Ng-Thow-Hing, C. M. Khursigara, N. Moiseeva, M. Allaire, and J. W. Coulton.** 2006. Structure of TonB in complex with FhuA, E. coli outer membrane receptor. Science **312:**1399-402.

102. **Perez-Losada, M., E. B. Browne, A. Madsen, T. Wirth, R. P. Viscidi, and K. A. Crandall.** 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol **6:**97-112.

103. **Pettigrew, M. M., B. Foxman, C. F. Marrs, and J. R. Gilsdorf.** 2002. Identification of the lipooligosaccharide biosynthesis gene *lic2B* as a putative virulence factor in strains of nontypeable *Haemophilus influenzae* that cause otitis media. Infect Immun **70:**3551-6.

104. **Porras, O., D. A. Caugant, B. Gray, T. Lagergard, B. R. Levin, and C. Svanborg-Eden.** 1986. Difference in structure between type b and nontypable *Haemophilus influenzae* populations. Infect Immun **53:**79-89.

105. **Poulsen, K., J. Reinholdt, and M. Kilian.** 1992. A comparative genetic study of serologically distinct *Haemophilus influenzae* type 1 immunoglobulin A1 proteases. J Bacteriol **174:**2913-21.

106. **Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson.** 2010. New approaches to population stratification in genome-wide association studies. Nat Rev Genet **11:**459-63.

107. **Pritchard, J. K., and P. Donnelly.** 2001. Case-control studies of association in structured or admixed populations. Theor Popul Biol **60:**227-37.

108. **Pritchard, J. K., M. Stephens, and P. Donnelly.** 2000. Inference of population structure using multilocus genotype data. Genetics **155:**945-59.

109. **Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly.** 2000. Association mapping in structured populations. Am J Hum Genet **67:**170-81.

110. **Pritchard, J. K., X. Wen, and D. Falush.** 2010. *structure* 2.3 documentation.

111. **R Development Core Team.** 2011. R: a language and environment for statistical computing. Version 2.13.0. R Foundation for Statistical Computing, Vienna.

112. **Redfield, R. J., W. A. Findlay, J. Bosse, J. S. Kroll, A. D. Cameron, and J. H. Nash.** 2006. Evolution of competence and DNA uptake specificity in the Pasteurellaceae. BMC Evol Biol **6:**82.

113. **Ridderberg, W., M. G. Fenger, and N. Norskov-Lauritsen.** 2010. Haemophilus influenzae may be untypable by the multilocus sequence typing scheme due to a complete deletion of the fucose operon. Journal of Medical Microbiology **59:**740-2.

114. **Riehle, M. M., W. M. Guelbeogo, A. Gneme, K. Eiglmeier, I. Holm, E. Bischoff, T. Garnier, G. M. Snyder, X. Li, K. Markianos, N. Sagnon, and K. D. Vernick.** 2011. A cryptic subgroup of Anopheles gambiae is highly susceptible to human malaria parasites. Science **331:**596-8.

115. **Rosenberg, N. A.** 2004. DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes **4:**2.

116. **Rosenberg, N. A., T. Burke, K. Elo, M. W. Feldman, P. J. Freidlin, M. A. Groenen, J. Hillel, A. Maki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, and S. Weigend.** 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics **159:**699-713.

117. **Rothman, K. J., and S. Greenland.** 1998. Modern epidemiology, 2nd ed. Lippincott-Raven, Philadelphia, PA.

118. **Roy, A., A. Kucukural, and Y. Zhang.** 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc **5:**725-38.

119. **Sacchi, C. T., D. Alber, P. Dull, E. A. Mothershed, A. M. Whitney, G. A. Barnett, T. Popovic, and L. W. Mayer.** 2005. High level of sequence diversity in the 16S rRNA genes of *Haemophilus influenzae* isolates is useful for molecular subtyping. J Clin Microbiol **43:**3734-42.

120. **Samuelson, A., A. Freijd, J. Jonasson, and A. A. Lindberg.** 1995. Turnover of nonencapsulated *Haemophilus influenzae* in the nasopharynges of otitis-prone children. J Clin Microbiol **33:**2027-31.

121. **Seale, T. W., D. J. Morton, P. W. Whitby, R. Wolf, S. D. Kosanke, T. M. VanWagoner, and T. L. Stull.** 2006. Complex role of hemoglobin and hemoglobin-haptoglobin binding proteins in *Haemophilus influenzae* virulence in the infant rat model of invasive infection. Infect Immun **74:**6213-25.

122. **Selander, R. K., D. A. Caugant, H. Ochman, J. M. Musser, M. N. Gilmour, and T. S. Whittam.** 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl Environ Microbiol **51:**873-84.

123. **Shen, K., P. Antalis, J. Gladitz, S. Sayeed, A. Ahmed, S. Yu, J. Hayes, S. Johnson, B. Dice, R. Dopico, R. Keefe, B. Janto, W. Chong, J. Goodwin, R. M. Wadowsky, G. Erdos, J. C. Post, G. D. Ehrlich, and F. Z. Hu.** 2005. Identification, distribution, and expression of novel genes in

10 clinical isolates of nontypeable *Haemophilus influenzae*. Infect Immun **73:**3479-91.

124. **Sheppard, S. K., F. Colles, J. Richardson, A. J. Cody, R. Elson, A. Lawson, G. Brick, R. Meldrum, C. L. Little, R. J. Owen, M. C. Maiden, and N. D. McCarthy.** 2010. Host association of Campylobacter genotypes transcends geographic variation. Appl Environ Microbiol **76:**5269-77.

125. **Sheppard, S. K., N. D. McCarthy, D. Falush, and M. C. Maiden.** 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. Science **320:**237-9.

126. **Smith, J. M., E. J. Feil, and N. H. Smith.** 2000. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays **22:**1115-22.

127. **Smith, J. M., N. H. Smith, M. O'Rourke, and B. G. Spratt.** 1993. How clonal are bacteria? Proc Natl Acad Sci U S A **90:**4384-8.

128. **St Sauver, J., C. F. Marrs, B. Foxman, P. Somsel, R. Madera, and J. R. Gilsdorf.** 2000. Risk factors for otitis media and carriage of multiple strains of *Haemophilus influenzae* and *Streptococcus pneumoniae*. Emerg Infect Dis **6:**622-30.

129. **Supply, P., R. M. Warren, A. L. Banuls, S. Lesjean, G. D. Van Der Spuy, L. A. Lewis, M. Tibayrenc, P. D. Van Helden, and C. Locht.** 2003. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. Mol Microbiol **47:**529-38.

130. **Tahtinen, P. A., M. K. Laine, P. Huovinen, J. Jalava, O. Ruuskanen, and A. Ruohola.** 2011. A placebo-controlled trial of antimicrobial treatment for acute otitis media. N Engl J Med **364:**116-26.

131. **Thomas, C. E., B. Olsen, and C. Elkins.** 1998. Cloning and characterization of tdhA, a locus encoding a TonB-dependent heme receptor from Haemophilus ducreyi. Infect Immun **66:**4254-62.

132. **Tong, Y., and M. Guo.** 2009. Bacterial heme-transport proteins and their heme-coordination modes. Arch Biochem Biophys **481:**1-15.

133. **Tristram, S., M. R. Jacobs, and P. C. Appelbaum.** 2007. Antimicrobial resistance in *Haemophilus influenzae*. Clin Microbiol Rev **20:**368-89.

134. **Trottier, S., K. Stenberg, and C. Svanborg-Eden.** 1989. Turnover of nontypable *Haemophilus influenzae* in the nasopharynges of healthy children. J Clin Microbiol **27:**2175-9.

135. **Turner, K. M., W. P. Hanage, C. Fraser, T. R. Connor, and B. G. Spratt.** 2007. Assessing the reliability of eBURST using simulated populations with known ancestry. BMC Microbiol **7:**30.

136. **Ukkonen, P., K. Varis, M. Jernfors, E. Herva, J. Jokinen, E. Ruokokoski, D. Zopf, and T. Kilpi.** 2000. Treatment of acute otitis media with an antiadhesive oligosaccharide: a randomised, double-blind, placebo-controlled trial. Lancet **356:**1398-402.

137. **Varrasso, D. A.** 2006. Otitis media: the need for a new paradigm in medical education. Pediatrics **118:**1731-3.

138. **Verdu, P., F. Austerlitz, A. Estoup, R. Vitalis, M. Georges, S. Thery, A. Froment, S. Le Bomin, A. Gessain, J. M. Hombert, L. Van der Veen, L. Quintana-Murci, S. Bahuchet, and E. Heyer.** 2009. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. Curr Biol **19:**312-8.

139. **Vitovski, S., K. T. Dunkin, A. J. Howard, and J. R. Sayers.** 2002. Nontypeable *Haemophilus influenzae* in carriage and disease: a difference in IgA1 protease activity levels. JAMA **287:**1699-705.

140. **Vitovski, S., and J. R. Sayers.** 2007. Relaxed cleavage specificity of an immunoglobulin A1 protease from *Neisseria meningitidis*. Infect Immun **75:**2875-85.

141. **Vos, M., and X. Didelot.** 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J **3:**199-208.

142. **Wald, E. R.** 2003. Acute otitis media: more trouble with the evidence. Pediatr Infect Dis J **22:**103-4.

143. **Wang, S., C. M. Lewis, M. Jakobsson, S. Ramachandran, N. Ray, G. Bedoya, W. Rojas, M. V. Parra, J. A. Molina, C. Gallo, G. Mazzotti, G. Poletti, K. Hill, A. M. Hurtado, D. Labuda, W. Klitz, R. Barrantes, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, E. Llop, F. Rothhammer, L. Excoffier, M. W. Feldman, N. A. Rosenberg, and A. Ruiz-Linares.** 2007. Genetic variation and population structure in native Americans. PLoS Genet **3:**e185.

144. **Whitby, P. W., T. W. Seale, T. M. Vanwagoner, D. J. Morton, and T. L. Stull.** 2009. The iron/heme regulated genes of Haemophilus influenzae: Comparative transcriptional profiling as a tool to define the species core modulon. BMC Genomics **10:**6.

145. **White, D. C., and S. Granick.** 1963. Hemin biosynthesis in *Haemophilus*. J Bacteriol **85:**842-50.

146. **Xie, J., P. C. Juliao, J. R. Gilsdorf, D. Ghosh, M. Patel, and C. F. Marrs.** 2006. Identification of new genetic regions more prevalent in nontypeable *Haemophilus influenzae* otitis media strains than in throat strains. J Clin Microbiol **44:**4316-25.

147. **Zhang, Y.** 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics **9:**40.

148. **Zhang, Y.** 2007. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins **69 Suppl 8:**108-17.